Exploration and Model of COVID-19 in U.S. Nursing Homes

Aretha Miller

12/10/2021

## Executive Summary

January 21, 2020, a man in his 30’s was the first confirmed case of COVID-19 in the U.S. [King 5 Staff, 2021](https://www.king5.com/article/news/health/coronavirus/first-us-coronavirus-case-one-year-later/281-ec4fb66b-c8c5-4d31-9662-7c8f5f0f5d20). About a month later, the second case of COVID-19 was confirmed in a teen boy about 10 miles away. Twenty four hours later, a man in his 50’s was the first confirmed U.S. death. He lived at the Lifecare Center of Kirkland, 25 miles away from the first confirmed case of COVID-19 [Google Maps](https://www.google.com/search?q=distance+of+providence+medical+center+everett+to+kirkland+lifecare&rlz=1C1GCEA_enUS893US893&sxsrf=AOaemvICbyCThbnnoQvCMDUTTcdROrcFPA%3A1639003462548&ei=RjWxYeaXII7O0PEPk86lwAQ&ved=0ahUKEwjm0q3Zo9X0AhUOJzQIHRNnCUgQ4dUDCA4&uact=5&oq=distance+of+providence+medical+center+everett+to+kirkland+lifecare&gs_lcp=Cgdnd3Mtd2l6EAM6BAgAEEc6BwgjELACECdKBAhBGABQhg1Y2SFggCtoAHACeACAAXSIAY8KkgEDOC42mAEAoAEByAEIwAEB&sclient=gws-wiz). That Long Term Care Facility became the epicenter of COVID-19 cases in the U.S., and ultimately 101 residents and 55 staff were confirmed positive for COVID-19 [King County Public Health, 2021](https://kingcounty.gov/depts/health/covid-19.aspx). These cases also resulted in 46 deaths [King 5 Staff, 2020](https://www.king5.com/article/news/health/coronavirus/vaccine/kirkland-nursing-home-with-first-us-coronavirus-outbreak-received-vaccines-monday/281-10e43c35-f3bb-4148-a1e1-1cdb8684de9c). Nowadays, The Centers for Medicare & Medicaid Services mandated Nursing Homes to report certain COVID-19 data to the [Centers for Disease Control and Prevention’s [CDC] National Healthcare Safety Network](file:///C:\Users\Chris\Documents\%22https:\data.cms.gov\covid-19\covid-19-nursing-home-data%22). Since we are almost two years into the pandemic, we speculated there may be a plethora of data to analyze.

We explored data that was submitted to the CDC with statistical techniques called Principal Component Analysis (PCA) and Clustering, and then modeled the data with another statistical method called Artificial Neural Network (ANN). We analyzed the proportions of total variance distribution, according to the principal component loadings, to assess the importance of the input variables. We explored whether there were a smaller number of groups into which the variables could be meaningfully clustered. We also queried which variables contributed the most to the predicted probability that residents may die from COVID-19, and how accurately we could predict that residents may die from COVID-19.

## Data Exploration

This data set contained 12,417 observations and 120 variables. We cleaned the data by encoding empty strings with NA and removed sparse columns with a sparsity threshold of 0.1. We eliminated missing values and excluded variables we planned to ignore. For instance, administrative columns like Federal Provider Numbers. Once cleaned, we included 11,640 rows, 40 columns with 15 continuous variables. All variables were right skewed with outliers (Figure 1), so we standardized and regularized the data set.

Chart, scatter chart

Description automatically generated

***Figure 1: Scatterplot of Right-Skewed COVID-19 Data set***

Principal Component Analysis: PCA is used to observe trends, outliers and decrease the dimension of a data matrix (James, Witten, Hastie & Tibshirani (2013). We standardized each variable to mean zero and standard deviation one, and then computed 15 distinct principal components (PCs). We then plotted the data points against each other to create a low dimensional view of the data. A biplot (Figure 2) can optimally represent distances between observations and relationships among variables (James et al., 2013).

Diagram

Description automatically generated

***Figure 2: Biplot showing distances among observations***

Next, we followed up with k-means clustering (Figure 3) to visualize four clusters of the data in a Cluster plot.

A picture containing chart

Description automatically generated

***Figure 3: Data in Four Clusters***

Hierarchial Clustering: We then proceeded to hierachial clustering with a goal to figure out whether the observations aggregated into a distinct number of groups. We plotted a dendrogram using complete linkage clustering with Euclidean distance as the similarity measure. We cut the dendrogram at a height(distance) of about 55 that yielded 4 clusters. Several observations may be seen at the bottom of the dendrogram, because the data set is huge.

Artificial Neural Network: ANN is a computational model that comprise of many processing components that accept inputs and deliver outputs based on predetermined activation functions (Nielsen, 2015). ANN was used to determine which variables contributed the most to the predicted probability that residents will die from COVID-19, as well as how accurately we can predict that residents will die from COVID-19. The NNET package from R programming was used in the experimentation with neural network to create a model and verify its accuracy.

A single layer of validation on the ANN was used to tune 15 predictor variables. Ten-fold cross validation with R’s caret package was used to build an ANN with five nodes to model “Residents Total COVID-19 Deaths” (response variable), as a function of all other variables in the data set. We determined the decay rate based on the value of lambda that optimized the Root Mean Square Error (RMSE) with an optimal lambda of one.

Diagram

Description automatically generated

***Figure 4: ANN with Five Nodes to model Residents Total COVID-19 Deaths***

Next, we conducted an outer layer of validation containing all the supervised learning models by using the training data, rather than the full data set. All variables were right-skewed, so standardization and regularization was used to re-scale all input and output variables before training the model. We plotted the Neural Network and performed a Garson plot to view variable importance. We also processed an Olden plot to view the direction and the association among the predictors and the response variable. Finally, we conducted a Lek profile to explore the relationship among the predictors and outcome variable.

Chart, bar chart

Description automatically generated

***Figure 5: Garson Plot showing most important predictors***

## Results

Initial exploration of the data revealed all 15 variables had vastly different means and the variances showed there were on average three times as many “Staff Total Confirmed COVID-19” as “Residents Total Confirmed COVID-19,” and more than six times as many “Staff Total COVID-19 Deaths” as “Residents Total COVID-19 deaths.”

The biplot (Figure 2) revealed the first two PCs of the data. We saw the concentration of most data points in the left lower quadrant with several outliers, even though standardization and regularization of the data was done. The first loading vector placed approximate equal weights on “Staff Weekly COVID-19 Deaths” and “Residents Weekly Admissions COVID-19,” but less weight on “Residents Weekly COVID-19 Deaths.” The second loading vector placed most of its weight on “Residents and Staff Weekly Confirmed COVID-19.” Overall, variables like “Staff Weekly COVID-19 deaths” and “Residents Weekly Admissions COVID-19” were close together and therefore correlated, while “Residents and Staff Weekly Confirmed COVID-19” were further out and may not be associated. Next, a computation of the proportion of variances showed the first PC explained 27% of the variance in the data and the second PC explained 41% of the variances. A screeplot was used to determine the number of factors to retain in an exploratory factor analysis. We saw an elbow in the plot at about two and this suggested there may be little benefit to examining more than two PCs.

The model summary from the ANN revealed the best concentration of data around the line of best fit (RMSE) was 4.63, how close the data was to the fitted line (R-squared or predictive ability of the model) was 84%, while the amount of error in our measurements (Mean Absolute Error) was 3.34. According to our Garson plot (Figure 5), the two most important predictors were “Residents Total All Deaths” and “Staff Total COVID-19 Deaths.”

The Olden Plot revealed that both predictors were positively associated with the outcome variable. We feel this model is sufficiently correct to be used for new predictions of selected data points. However, model accuracy may be boosted by addressing outliers, by removing, transforming, or imputing them [Ray, 2015](https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/). Furthermore, since the data is non-linear, we may also choose a different machine learning algorithm, like Random Forest. Finally, the Lek profile displayed the relationships among the variables.

Calendar

Description automatically generated

***Figure 6: Lek Profile Displaying Relationships Among Variables***

The first variable in the Lek Profile was “Residents Total Admissions COVID-19” and we saw a horizontal line at about the 20th percentile. Most of the other variables were bunched to the left side of their plots, while others were difficult to visualize. Nevertheless, visualizing a neural network with plotnet may be impractical for large models (Beck, 2018), but plots like the Garson algorithm defined the relationships among variables by displaying their importance.

## Discussion

Some ways in which the retained data differed from the original data set were the exclusion of columns that had binary categorical data. For instance, all columns that addressed personal protective equipment (PPE) were excluded, because once the data were cleaned and “NA’s” removed, the columns were too sparse for meaningful analysis.

Both of the two most important predictors and their relationships with the response variable were surprising. “Staff Total COVID-19 Deaths” was the second most important variable. This created questions like whether the staff were tested regularly for COVID-19, whether they had access to and wore PPE, whether they received training on how to manage the spread of COVID-19, or whether the staff were fully vaccinated.

Our target audience is The American Healthcare Association and the National Center for Assisted Living (AHCA/NCAL) which is the largest organization in the U.S. representing long term and post-acute care providers. As of 2021, the CDC mandated Nursing Homes to report COVID-19 vaccination of staff and residents (U.S. Department of Health & Human Services, 2021). The target audience may improve the response variable by reporting the vaccination status of residents and staff as soon as possible. It is plausible that the relationship among the predictors and response variable could be causal, since Healthcare Personnel could shed the COVID virus to residents unknowingly.

## Conclusion

As we remain in an active pandemic, COVID-19 is fluid and ever changing. In this study, we analyzed a public data set using unsupervised and supervised machine learning techniques. We explored a new and evolving data set and demonstrated the importance of the input variables, according to the variance distribution of their PC loadings. Next, we illustrated that variables can be meaningfully clustered into a smaller number of groups and also displayed which variables contributed the most to the predicted probability that residents may die from COVID-19. The social implications from this research is to help future researchers to build on this study and the AHCA/NCAL to improve the lives of their vulnerable population.

## References

Beck, M.W. (2018). NeuralNetTools: Visualization and analysis tools for neural networks. Journal of Statistical Software,85(11). <doi:10.18637/jss.v085.i11>

Centers for Disease Control and Prevention. (2021, November 9). LTCF COVID-19 Module. Retrieved from <https://www.cdc.gov/nhsn/ltc/covid19/index.html>

Centers for Medicare and Medicaid Services. (2021, November 21). COVID-19 Nursing Home Data. Retrieved from <https://data.cms.gov/covid-19/covid-19-nursing-home-data>

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. New York: Springer

King County Public Health. (2021). COVID-19 Information and Resources. Retrieved from <https://kingcounty.gov/depts/health/covid-19.aspx>

King 5 Staff. (2021, January 21). Snohomish County nurse who treated 1st confirmed US case of COVID-19 reflects on the past year. Retrieved from <https://www.king5.com/article/news/health/coronavirus/first-us-coronavirus-case-one-year-later/281-ec4fb66b-c8c5-4d31-9662-7c8f5f0f5d20>

King 5 Staff. (2020, December 28). Kirkland Nursing Home with first U.S. Coronavirus Received Vaccines Monday. Retrieved from <https://www.king5.com/article/news/health/coronavirus/vaccine/kirkland-nursing-home-with-first-us-coronavirus-outbreak-received-vaccines-monday/281-10e43c35-f3bb-4148-a1e1-1cdb8684de9c>

Nielsen, M.A.(2015). Neural networks and deep learning. Determination Press. Retrieved from <http://neuralnetworksanddeeplearning.com/>

Ray, S. (2015). Eight proven ways for improving the “Accuracy” of a machine learning model. Retrieved from <https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/>

U.S. Department of Health & Human Services. (2021). Weekly HCP & Resident COVID-19 Vaccination. Retrieved from <https://www.cdc.gov/nhsn/ltc/weekly-covid-vac/index.html>