Direct Multi-Step Forecasting with Multiple Time Series Using XGBoost: Projecting COVID-19 Positive Hospitalization Census for a Southern Idaho Health System

Drake Anshutz, MPH[1], Andrew Crisp[2], James Ford, MS[3], Onur Torusoglu, MBA[4], Justin Smith, PhD[5]

Digital and Analytics: Advanced Analytics, St. Luke's Health System, Boise, ID[1,2,3,5]
Digital and Analytics, St. Luke's Health System, Boise, ID[4]
anshutzd@slhs.org[1], smitjust@slhs.org[5]

Abstract

Forecasting hospitalization census for the novel COVID-19 virus is a challenging task for numerous reasons including many unknowns, limited historical data, and other issues related to model misspecification. Most modeling techniques aimed at predicting hospitalization census for respiratory epidemics often create contradictory projections from a wide variety of scenarios. This often creates massive confidence intervals for projections as most models are based on manually adjusted assumptions which ultimately provide inconsistent, unreliable results. This case-study introduces a machine learning approach that helps overcome limited historical data while adjusting for model misspecification and creating consistent, easily understood results. This model has been deployed and automated with daily updates within a large health system for executive use and is reliably forecasting a one-month projection within an acceptable margin of error as determined by executive leadership.

COVID-19 Background

COVID-19 is a highly infectious novel disease that was declared a pandemic by the World Health Organization on March 11th, 2020 (Meehan et al, 2020; Baloch, 2020; Roda et al 2020). Modeling for the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which causes the coronavirus disease (COVID-19) has challenged data scientists and statisticians across the world for numerous reasons, to include but not limited to a novel disease with many unknowns, limited historical data, and policy shifts affecting the trajectory of the disease (Roda et al, 2020; Wang, 2020). Accurately predicting an outcome of COVID-19 positive patient hospitalization census has become an extreme task in this regard as rapidly changing policy enactments, shifts in human behavior, and other events such as masking ordinances and masking compliance strongly influence an outcome such as hospitalization census (Roda et al, 2020; Eikenberry, 2020). A strong modeling system that accounts for such factors is needed to inform future policies such as organizational decision making at an executive level (McBryde, 2020).

Early indications of COVID-19 hospitalization census are known to display cyclic tendencies due to influxes and outflows of patients over an extended time-period (Roda et al, 2020; Fiore et al, 2020). Traditional epidemiologic forecasts used to study disease behavior, such as basic SEIR (Susceptible Exposed Infectious Recovered) models, generally do not account for these cyclic tendencies and assume a smooth, normally distributed projection generally for a large scale population which creates discrepancies for localized clinical projections such as a singular hospital (Anirudh, 2020; Chen et al, 2020). If a traditional model is aimed at adjusting for an influx or outflow of patients for a particular population, the modeler can manually adjust certain assumptions, such as transmissibility, susceptible patients, and hospitalization rate within the model but such assumptions are highly prone to inaccuracies in future projections (Eksin, Paarporn, Weitz, 2019; Huppert and Katrield. 2013). This not only creates inaccuracies in the distribution of patients over time but competing assumptions within various models often contradict each other resulting in unknown accuracy across modeling techniques (Wang et al, 2016).

Model Selection

Direct Multi-Step Forecasting with Multiple Time Series is a method that directly projects a continuous outcome specific to each time-step (Redell, 2020; Guillaume and Chevillon, 2005; Taieb and Atiya, 2016). Other forecasting methodologies generally predict in a recursive nature to follow the trajectory of a projection (Guillaume and Chevillon, 2005). Direct Forecasting specifies a target date and uses lagged variables and dynamic predictive features across forecasting horizons to predict certain outcomes at a designated point in time (Redell, 2020). Strengths to this forecasting methodology include robustness to policy enactments, the ability to train on less data eliminating the cold-start problem, addresses a concept called model misspecification within a reasonable time-period and provides consistent model results (Redell, 2020; Guillaume and Chevillon, 2005; Chen et al, 2020).

XGBoost is an ensemble learning method that follows a gradient boosted tree format and is frequently used in other healthcare machine learning models because of its performance (Chen et al, 2020; Xu et al, 2019; Liu et al, 2018). This algorithm is run for a continuous outcome (XGBoost for regression) and is particularly useful in analyzing variable importance eliminating the "black-box" issue common among other machine learning techniques (Chen and Carlos, 2016).

Sample and Data Sources

The dataset used is an automated blending of two data sources. The first is a dataset updated daily from an internal data management team that provides a refreshed dataset on hospitalization census, including admissions and discharges from prior mid-night census. The patient volumes for this model are refreshed daily and the groupings for hospitalization census include all COVID-19 positive patients within four separate hospitals and intensive care unit census split between two regions (Two intensive care units for one region and four for the other region).

An additional data source is collected from the Johns Hopkins University Center for Systems Science and Engineering using the R package covid19.analytics. This dataset measures the total of new positive COVID-19 cases as reported by the State of Idaho (Ponce, 2020). New positive COVID-19 cases are defined by county and allocated to relevant hospitals within the determined service region of each hospital (i.e. if a county is located within a hospitals service region, the positive cases from that county as reported by the covid19.analytics package will be attributed to that particular hospital). State reporting was ultimately chosen to depict new cases of COVID-19 as internal data sources for testing remain unstandardized in terms of shifting market shares for testing across competing health systems. Counties and

hospital locations are specifically withheld from this report to ensure patient privacy.

Additionally, Idaho State Reopening Phases (i.e. mandates that require certain businesses such as bars and restaurant's to close or remain partially open for a specific period of time) were incorporated as well as holiday weekends and both were included as binary dynamic features (State of Idaho, 2020). Day numbers and week numbers were also included as dynamic features.

The final dataset ultimately includes two separate types of features or predictor variables. The continuous lagged features are total hospitalization census, total discharges, total admissions, and total positive cases. The dynamic features included are Idaho State Reopening Phases, day number, week number, and holiday weekends.

HIPAA Compliance

All patient data has been aggregated and anonymously displayed within de-identified hospitals. No individual patient data was used for analysis and the project followed the Privacy Rule as stated by the Health Insurance Portability and Accountability Act for de-identifying all 18 elements of identity (US Department of Health and Human Services, 2020). In addition, no demographics were reported to ensure patient privacy.

Analyses

R was used for all analyses (R Core Team, 2013). Direct Multi-Step Forecasting with Multiple Time Series using the Machine Learning Algorithm XGBoost was employed as the model to forecast hospitalization mid-night census and intensive care unit mid-night census. The R package used for analysis was **forecastML** (Redell, 2020). The parameters used for the two outcomes of hospitalization census and intensive care unit census are as follows; Lookback: 140 days for both Hospital and ICU, Horizons: 1, 14, and 30-days for both Hospital and ICU, Frequency: 1-day for both Hospital and ICU.

XGBoost was employed to project the value per time-step. Regression for squared error was chosen as the objective. Default settings within the XGBoost parameters and function were employed. The validation metric used is Mean Absolute Error. Prediction confidence intervals of +/- 2 were used.

Validation Process

The validation process follows a nested cross-validation setup (Redell, 2020; Bergmeir, Hyndman, Koo, 2017). Validation is examined by extracting validation windows and analyzes model performance with those validation windows

across selected horizons. The amount of time (in this report, days) within the validation window ultimately serves as the testing set for model performance. So, if a validation window of 9 days is selected, 9 projections will be created for each of those days within the validation window. The differences between the projection and actual will serve as the validation metric and this report has selected mean absolute error to represent the metric for model performance. This report evaluates the validation window selected for three separate horizons (1-day, 14-day, and 30-day) and combines the results from all validation windows and horizons to provide a global mean absolute error. The report also examines validation windows across time, separated by model horizons to depict accuracy of the model horizon performance throughout time.

Results

The global mean absolute error (GMAE) for both models (Hospitalization and Intensive Care Units) at the window and skip ratio of 9-day windows and 21-day skips was used to assess model performance. Validation was also analyzed over-time via mean absolute error (MAE) to assess the accuracy of the model's performance at the same 9-day window and 21-day skip ratio.

Hospital	Global Mean Absolute Error
Hospital 1	1.43
Hospital 2	2.37
Hospital 3	2.00
Hospital 4	1.22
ICU 1	1.21
ICU 2	1.08

Table 1: Global Mean Absolute Error across Hospitals and Intensive Care Units (ICU)

Global Mean Absolute Error
The forecasting GMAE stayed consistent across Hospitals and Intensive Care Units with highest GMAE found at 2.37 which is displayed in Table 1. Other validation window and skip ratios were evaluated and the highest GMAE found was at a 30-day window and 30-skip with a value of 3.55. The ranges for the 30-day window and 30-day skip were 1.26 – 3.55 across both the hospitalization and intensive care unit models.
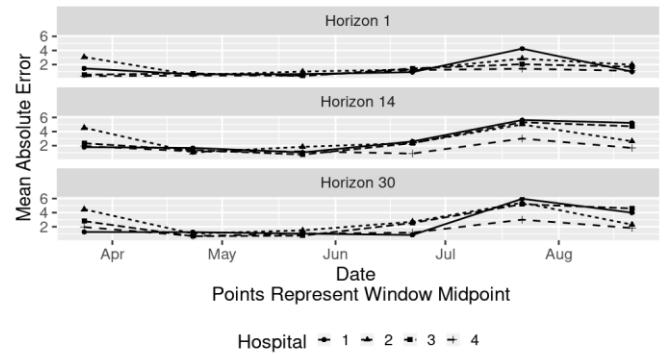

Figure 1: Forecasting Error Across Validation Windows for a 9-Day Window and 21-Day Skip for Hospitalizations
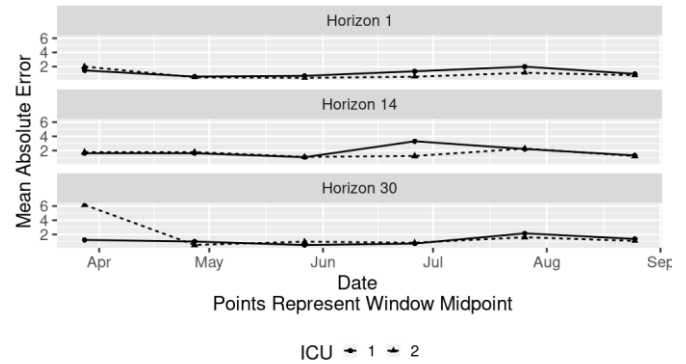

Figure 2: Forecasting Error Across Validation Windows for a 9-Day Window and 21-Day Skip for Intensive Care Units (ICU)

Mean Absolute Error Across Windows and Horizons
The forecasting MAE displayed in Figures 1 and 2 display the fluctuations across forecasting windows which remain consistent across time with a maximum MAE of 5.96 for Hospitalizations and 6.17 for ICU. The other validation ratio of 30-day windows and 30-day skips found a maximum MAE of 9.61 with a range of 0.63 – 9.61 for Hospitalizations. The same 30-day window and 30-day skip for ICU's found a maximum MAE of 5.38 with a range of 0.90 – 5.38.

Variable Importance Assessment
Variable importance was analyzed and assessed daily. Most variable importance gain was identified in total hospitalizations lag variables, discharge and admissions lagged variables and positive testing lagged variables. The consensus of variable importance indicates the model was dependent on which model horizon the model was assessing. For instance, the one-day projection was relatively dependent on the previous weeks' total hospitalizations, admissions, discharges, and positive testing. However, a thirty-day forecast would utilize much different variables that were dependent on longer term historical data. Additionally, Idaho State Phases were rarely identified in the model importance. However, a

holiday weekend would indicate a high value in gain if a holiday were within a horizon.

Discussion

This paper describes a novel approach toward predicting COVID-19 positive patient hospitalization census that has not been seen in recent literature as of current date. The model is currently in use and projecting within a reasonable mean absolute error and most training validations were performed with less than 6-months' worth of historical data. The value of this predictive model also includes the introduction of future policy decisions of which can be automated to be included in future iterations of the model (Redell, 2020).

The models forecast also adjusts for the forecasting horizons for the final projection delivered to executive leaders. This ultimately lowers error across forecasting horizons as the model uses the one-day forecast for the first prediction in the final projection, the fourteen-day forecast for days 2:14, and the thirty-day forecast for days 15:30 (Selim et al, 2020; Taieb et al, 2020). After evaluating the results of the forecasting error validation windows, this will start at generally the strongest model projection and lead into slightly weaker forecasting projections.

Variable importance depicted within the model help illustrate why the model is predicting what it is predicting and uncovering the "black-box" of the algorithm. Understanding variable importance often allows a decision-maker to understand potential interventions that could limit a negative outcome (i.e. if a decision-maker is able to understand a future negative-outcome relationship with a modifiable predictor variable, the decision-maker may be able to adjust that predictor variable to establish a potential positive outcome). Unfortunately, most lagged features within this model will not be able to be influenced as the occurrence has already happened. However, a dynamic feature such as State Reopening Phases may be altered if a decision-maker chooses to assess and enact potential interventions.

Limitations

This model is still evolving and being tuned during a highly unpredictable time-period. This methodology can also be prone to issues described as a broken-curve and is slightly prone to overfitting historical data (Selim et al, 2020).

At this point, the model is strongly recommended to assess the trajectory of hospitalization census and not be used for individual day decision-making. The current models' projection, in comparison to actuals, tends to project influxes and outflows consistently with actual data. However, the specific date of the influx or outflow tends to occur within one to three days of the actual result. An actual influx of hospitalization census may occur on a Wednesday, but

the projection may predict the influx would happen on the prior Tuesday or the following Friday.

Strengths

The model is currently validated within reasonable margin of error as determined by executive leadership for business decision making. Given most time-series models require an abundant data source with multiple years to project future outcomes, this model can assess a future projection with a limited dataset (Redell, 2020). The addition of future policies may also be incorporated to help predict future interactions within the model (Redell, 2020). The model is also intended to gain strength in predictions as more data is incorporated into the models. Added robustness towards model misspecification is an additional strength within direct forecasting (Marcellino, Stock, Watson. 2006). Other strengths include that the model is currently refreshed daily to provide more accurate results as the future dates occur. A robust internal dataset is also an asset as most organizations outside of health systems are not granted access to live data sources with a large quantity of potential predictive features. Ultimately, the model can help determine future trends to assist executive leadership with resource utilization across a health system.

References

Anirudh A. 2020. Mathematical modeling and the transmission dynamics in predicting the Covid-19 - What next in combating the pandemic. *Infectious Disease Modelling*, *5*, 366–374.

Baloch, S., Baloch, M. A., Zheng, T., & Pei, X. 2020. The Coronavirus Disease 2019 (COVID-19) Pandemic. *The Tohoku journal of experimental medicine*, *250*(4), 271–278.

Ben Taieb, S., & Atiya, A. F. 2016. A Bias and Variance Analysis for Multistep-Ahead Time Series Forecasting. *IEEE transactions on neural networks and learning systems*, *27*(1), 62–76.

Bergmeir, C. Hyndman, R. & Koo, B. 2017. *A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Predictions.* Elsevier. robjhyndman.com/papers/cv-wp.pdf

Chen, S., Robinson, P., Janies, D., & Dulin, M. 2020. Four Challenges Associated With Current Mathematical Modeling Paradigm of Infectious Diseases and Call for a Shift. *Open forum infectious diseases*, *7*(8), ofaa333.

Chen, T., Carlos, G. 2016. XGBoost: A Scalable Tree Boosting System. *Association for Computing machinery, KDD '16*, 785-794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., & Li, Y. 2020. xgboost: Extreme Gradient Boosting. R package version 1.1.1.1. https://CRAN.R-project.org/package=xgboost

Eikenberry, S. E., Mancuso, M., Iboi, E., Phan, T., Eikenberry, K., Kuang, Y., Kostelich, E., & Gumel, A. B. 2020. To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infectious Disease Modelling*, 5, 293–308.

Eksin, C., Paarporn, K., & Weitz, J. S. 2019. Systematic biases in disease forecasting - The role of behavior change. *Epidemics*, 27, 96–105.

Fiore, V. G., DeFelice, N., Glicksberg, B. S., Perl, O., Shuster, A., Kulkarni, K., O'Brien, M., Pisauro, M. A., Chung, D., & Gu, X. 2020. Containment of future waves of COVID-19: simulating the impact of different policies and testing capacities for contact tracing, testing, and isolation. *medRxiv : the preprint server for health sciences*, 2020.06.05.20123372.

Guillaume, & Chevillon. 2005. DIRECT MULTI-STEP ESTIMATION AND FORECASTING N ° 2005-10 Juillet 2005. https://doi.org/10.1111/j.1467-6419.2007.00518.x

Huppert, A., & Katriel, G. 2013. Mathematical modelling and prediction in infectious disease epidemiology. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 19(11), 999–1005.

Liu, L., Yu, Y., Fei, Z., Li, M., Wu, F. X., Li, H. D., Pan, Y., & Wang, J. 2018. An interpretable boosting model to predict side effects of analgesics for osteoarthritis. *BMC systems biology*, 12(Suppl 6), 105.

Marcellino, M., Stock, J., & Watson, M. 2006. A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series. *Journal of Econometrics. 2006, 135:*499-526.

McBryde, E. S., Meehan, M. T., Adegboye, O. A., Adekunle, A. I., Caldwell, J. M., Pak, A., Rojas, D. P., Williams, B. M., & Trauer, J. M. 2020. Role of modelling in COVID-19 policy development. *Paediatric respiratory reviews*, 35, 57–60.

Meehan, M. T., Rojas, D. P., Adekunle, A. I., Adegboye, O. A., Caldwell, J. M., Turek, E., Williams, B. M., Marais, B. J., Trauer, J. M., & McBryde, E. S. 2020. Modelling insights into the COVID-19 pandemic. *Paediatric respiratory reviews*, 35, 64–69.

Ponce M. 2020. covid19.analytics: Load and Analyze Live Data from the CoViD-19 Pandemic.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org/. R package version 1.1.1. https://CRAN.R-project.org/package=covid19.analytics.

Redell N. 2020. forecastML: Time Series Forecasting with Machine Learning Methods. R package version 0.9.0. https://CRAN.R-project.org/package=forecastML

Roda, W. C., Varughese, M. B., Han, D., & Li, M. Y. 2020. Why is it difficult to accurately predict the COVID-19 epidemic? *Infectious Disease Modelling*, 5, 271–281.

Selim, M., Zhou, R., Feng, W., & Alam, O. 2020. Reducing error propagation for long term energy forecasting using multivariate prediction. *EPiC Series in Computing. 69,*161-169.

State of Idaho. 2020. Idaho Rebound: Path to Prosperity. Idaho Official Government Website. Retrieved from https://rebound.idaho.gov/stages-of-reopening/

Taieb, S, B., Bontempi, G. Atiya, A., & Sorjamaa, A. 2011. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Systems with Applications.* URL: http://souhaib-bentaieb.com/pdf/2012_esa_review.pdf

US Department of Health and Human Services. 2020. HIPAA Privacy Rule: Information for Researchers. Retrieved from https://privacyruleandresearch.nih.gov/

Wang J. 2020. Mathematical models for COVID-19: applications, limitations, and potentials. *Journal of public health and emergency*, 4, 9.

Wang, W. Liu, Q. Zhong, L. Tang, M. Gao, H. & Stanly, E. 2016. Predicting the epidemic threshold of the susceptible-infected-recovered model. *Sci Rep, 6*(24676), 2016.

Xu, Y., Yang, X., Huang, H., Peng, C., Ge, Y., Wu, H., Wang, J., Xiong, G., & Yi, Y. 2019. Extreme Gradient Boosting Model Has a Better Performance in Predicting the Risk of 90-Day Readmissions in Patients with Ischaemic Stroke. *Journal of stroke and cerebrovascular diseases: the official journal of National Stroke Association*, 28(12), 104441