Predicting New COVID-19 Cases in the US: A Comparison of a Compartmental Model (SIR) and XGBoost Machine Learning Model

Basil Okola

February 11, 2026

Background: Figuring out the number of new COVID-19 cases was key in aligning interventions to hotspots. This included ensuring there were enough beds for admissions, key protection gear like face masks, oxygen supply and intensive care unit space. A lot of mechanistic models were build borrowing from the traditional SIR model. Additionally, researchers explored machine learning techniques to predict the course of the infections.

Methods: In this report, we compare the performance of a SIR model to an XGBOOST prediction model in predicting new cases in each of the American states, 3 years after the pandemic.

Results: The XGBOOST ML approach demonstrated robust goodness-of-fit, effectively capturing the non-linear peak dynamics across varying geographic replicates.

1 INTRODUCTION

We chose to use data on COVID-19 cases assembled by a team from The London School of Health & Tropical Medicine who built a global COVID-19 tracker shiny app. This was purely due to convenience and the short turn around time for the assessment as there was an already ongoing personal project that modularized parts of the said app into a golem framework and so it was easier adding a prediction modelling capabilities rather than starting a new project. The COVID19Dash app can be installed from github and run seamlessly. For predictions look at the `Prediction Model` tab.

2 METHODS

2.1 Data

We used `cv_states` data that is part of the datasets in the `COVID19Dash` package. It compiles daily COVID-19 cases in each of the states in the US. It also comes with limited spatial support (longitudes & latitudes). We chose `new_cases` as the outcome variable and generated rolling means for 7, 14, and 30 days as predictors in addition to the frequency of each state in the data to avoid having to encode the categorical variable.

2.2 Models

We fitted a SIR model - a compartmental mechanistic framework that simulates the spread of infectious diseases by transitioning a population through susceptible, infectious, and recovered states using differential equations. It relies on the interaction between the transmission rate β and the recovery rate γ to determine the velocity and peak of an epidemic curve[1].

2.2.1 SIR framework

The mechanistic sir model assumes a fixed population N where individuals transition between susceptible (S), infectious (I), and recovered (R) compartments. the dynamics are governed by the following system of ordinary differential equations, which were implemented using the `desolve` package in our workflow:

$$\begin{aligned}\frac{dS}{dt} &= -\beta\frac{SI}{N} \\ \frac{dI}{dt} &= \beta\frac{SI}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

where β represents the effective transmission rate and γ denotes the removal or recovery rate. individuals are assumed to move from the susceptible to the infectious compartment at a rate proportional to the product of S and I.

A critical threshold in this modeling framework is the basic reproduction number (R_0), which defines the average number of secondary infections produced by a single infected individual in a completely susceptible population. in our analysis, we calculated R_0 using the optimized parameters from each state:

$$R_0 = \frac{\beta}{\gamma}$$

An epidemic persists only when $R_0 > 1$. our model fitting used an optimization routine to minimize the residuals between the predicted $I(t)$ compartment and the observed case ratios across the 50 american states.

2.2.2 XGBOOST model

We also fitted an XGBOOST model to predict the new cases. XGBoost, or Extreme Gradient Boosting, is an ensemble decision-tree algorithm, like random forest regressions, but able to model more complex interactions due to its ability to boost individual trees and does not rely on a single tree. It uses a scalable tree-boosting system to optimize predictions[2].

2.3 Missing data

We did not encounter missing data. However, depending on the objective of the research, there are a number of pathways available. In a purely ML prediction setting, researchers apply simple methods like mean/median for numeric variables or mode for categorical variables. In a purely inferential task, multiple imputations are employed under missing at random (MAR) assumption and a sensitivity analysis done to verify that the imputation method is not driving model decisions.

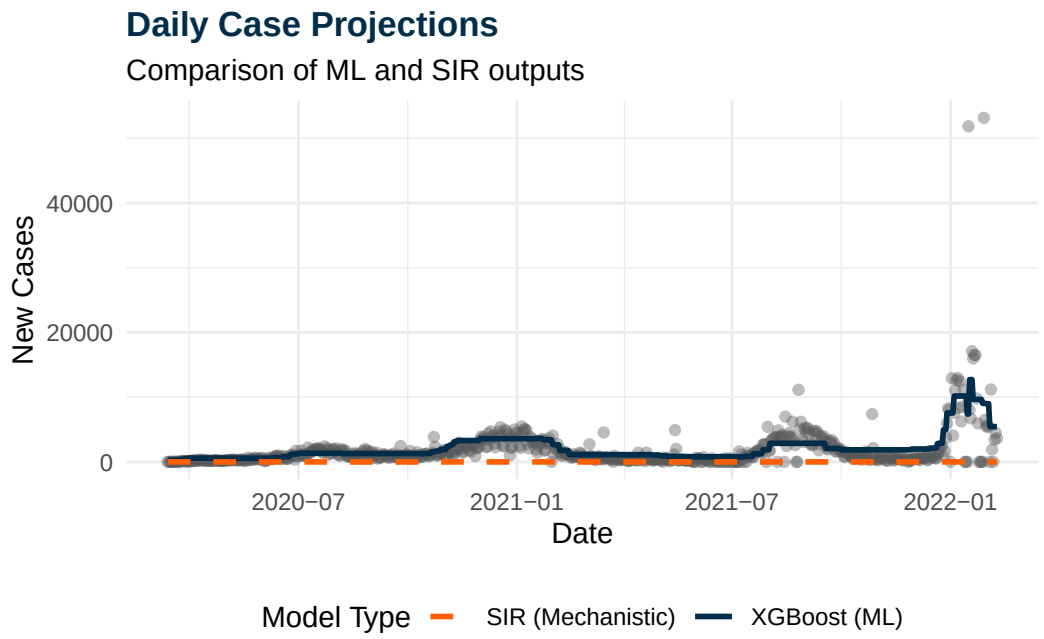
2.4 Accounting for spatial and tempororal structure

We did not work with a hierarchical spatial and temporal variation. However, one approach is through Hierarchical Bayesian models. In R this can be accomplished through R-INLA[3]. This involves decomposing observed variance into structured and unstructured random effects across multiple levels, using neighboring data to inform estimates in sparse regions.

3 RESULTS

3.1 Model comparisons

The XGBOOST performed better than the simple SIR model in predicting new cases.



4 DISCUSSIONS

We have build two models and compared their performance using visual plots and other metrics logged to the shiny app. We did not however undertake a simulation due to time constraints. We

consulted AI on figuring out bugs from XGBOOST which we found to be a challenge running in R environment.

5 REPRODUCIBILITY

Install the COVID19Dash using `devtools::install_github("bokola/COVID19Dash")` command. You can then run the app interactively by calling `run_app()`. To reproduce the contents of this document, ensure the other packages called in `setup` chunk are installed.

REFERENCE

1 Ross R. An application of the theory of probabilities to the study of a priori pathometry.—part i. *Proceedings of the Royal Society of London Series A, Containing papers of a mathematical and physical character* 1916; **92**: 204–30.

2 Chen T. XGBoost: A scalable tree boosting system. *Cornell University* 2016.

3 Moraga P. Geospatial health data: Modeling and visualization with r-INLA and shiny. Chapman; Hall/CRC, 2019.