Using GLMs and Machine Learning Through Tidymodels to Predict Public Health Scenarios

Matthew Krogman

September 9, 2025

Abstract

In recent years, air quality has become a more common topic of discussion among scientists and common people concerned about weather trends, especially during the summer months. Environmental change heightens the necessity of proper reaction and preparation for weather events or other phenomena that worsen air quality to the point of public health risk. In this report, I use generalized linear models (GLMs) as well as machine learning techniques to develop models that predict whether or not conditions are appropriate for a lockdown in an area, as well as the number of hospital visitors a hospital can expect on a single day. What I found in the GLMs confirms that there is interactive potential among the "six common pollutants", but only when not accounting for other environmental factors, like industrial activity. Machine learning assisted in creating useful predictions for the lockdown status and the number of hospital admissions on a day based on many factors, although the average error rate is a bit larger than expected. With that in mind, though, the machine learning models offer a valuable reference point for further research on the topic of machine learning for the sake of public health.

Introduction

As climate change continues to progress, and the population of major urban and industrial centers increases, more attention has been given to the quality of the air and the danger of breathing polluted air. Although emissions of harmful chemicals in the United States have reduced, thanks in part to the Clean Air Act, the American Lung Association claims that, as of 2024, 131 million Americans live in areas with unhealthy levels of air pollution.[1] This still-existent trend of more frequent and longer-lasting wildfires.[2] This environmental trend prompts the necessity to create models so that we may be better prepared to handle worsening air quality in the future. With data on pollutants and certain demographic data, it should be possible to prepare for lockdowns based on air quality and predict the daily number of hospital admissions in a specific location.

Data

The data to be used in this report is from a synthetic dataset created for Kaggle by Khushi Yadav, using a license from MIT. It is a time series spanning from the beginning of 2020 to the end of 2021. A large number of quantitative variables are included, such as AQI, industrial activity, and measures of the "six common pollutants" ($PM_{2.5}$, PM_{10}, O_3, CO, SO_2, and NO_2). Qualitative variables like the region and whether or

[1] "2024 'State of the Air' Report Reveals Most 'Hazardous' Air Quality Days in 25 Years." American Lung Association, Saint Paul, United States. April 24, 2024. https://www.lung.org/media/press-releases/sota-2024.

[2] Rickly, P. S et al.: "Emission factors and evolution of SO2 measured from biomass burning in wildfires and agricultural fires", Atmos. *Chem. Phys., 22, 15603–15620*, https://doi.org/10.5194/acp-22-15603-2022, 2022. 15604.

not a lockdown has been issued are also included. Throughout this report, all variables in the dataset will be used in multiple contexts to create effective prediction models for future use.

Though the data may be synthetic, the results are still true to the real world; it was generated using public domain resources on realistic statistical distributions for accuracy, while protecting privacy by limiting real-world attachment to individual data points.

The Appeal and Drawbacks of GLMs

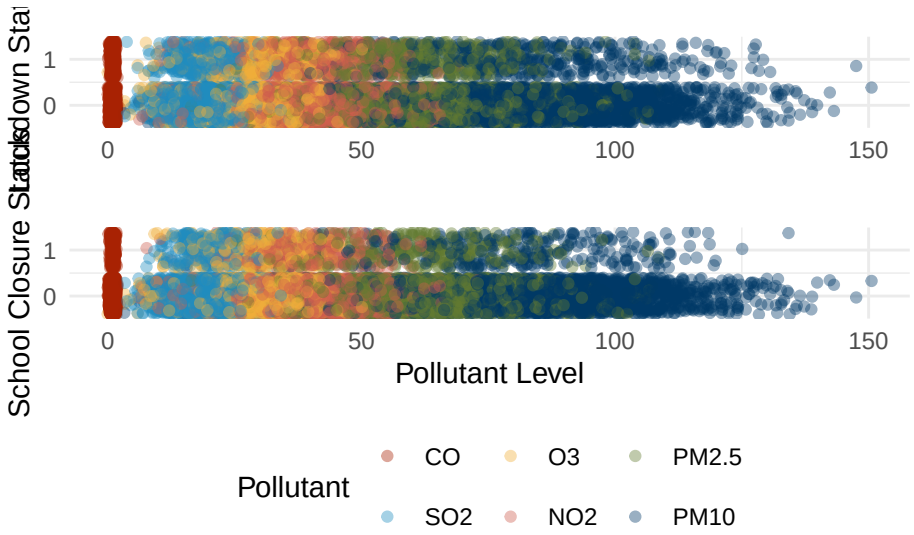
The state of the data provides a good opportunity to consider using generalized linear models: the response variables we are interested in lockdown state (binary) and hospital admissions (count) violate the assumptions of standard linear modeling. Creating a GLM is simple, but they are easy to overfit. A somewhat explanatory model for lockdown conditions is as follows:

Y ~indep. Bernoulli(p)
$log(\frac{p}{1-p}) = -2.085 - 0.017PM_{2.5} - 0.043SO_2 + 2.732CO + 0.002(PM_{2.5} * SO_2) - 0.021(PM_{2.5} * CO) - 0.082(SO_2 * CO)$

$SE_{PM_{2.5}} = 0.02, p = .38$
$SE_{SO_2} = 0.057, p = .45$
$SE_{CO} = 0.953, p = .004$
$SE_{PM_{2.5}*SO_2} = 0.001, p = .006$
$SE_{PM_{2.5}*CO} = 0.011, p = .07$
$SE_{SO_2*CO} = 0.035, p = .02$

There is quite a bit of doubt to be cast on this model, but it does also pique interest for further consideration. Indeed, there are a number of terms that provide strong evidence for an effect on lockdown status ($p < .05$). Of note is the interaction term between $PM_{2.5}$ and SO_2. This can be directly related to environmental factors, like wildfires. Wildfire smoke releases large amounts of $PM_{2.5}$, which in turn contains sulfate (SO_4),[3] which can become sulfur dioxide under circumstances like wildfire.

The possibility of an interaction certainly seems possible based on previous research, but it is possible to see this with a plot that puts this problem in the context of our data scenario. See below:



[3]Ricky et al. 15604.

The main takeaway from this plot is that in both cases of lockdowns and school closures, the pollutant level remains in the same region along the x-axis, regardless of if the observation was part of a day where there was a lockdown or not. The fact that the regions are the same suggests that there is an interactive effect between one or multiple pollutants that has an affect on air quality, such that a lockdown or school closure would be called. This is corroborated by the model as stated. Though we are not interested in school closures as a response variable for this report, it works with lockdown status to arouse suspicion about interaction terms in this scenario.

The other response variable we will focus on is the number hospital admissions in a day, made by combining variables relating to hospital resources.[4] That model is as such:

Y ~Poisson(λ)
$log(Y = y) = 3.519 - 0.001PM_{10} + 0.006SO_2 + 0.06CO + 0.001(PM_{10} * SO_2) + 0.0002(PM_{10} * CO) - 0.007(SO_2 * CO)$

$SE_{PM_{10}} = 0.0008, p = .07$
$SE_{SO_2} = 0.003, p = .08$
$SE_{CO} = 0.058, p = .30$
$SE_{PM_{10}*SO_2} = 0.00003, p = .43$
$SE_{PM_{10}*CO} = 0.0005, p = .045$
$SE_{SO_2*CO} = 0.002, p = .0006$

Much like the previous model, a couple of predictors show strong evidence of nonzero impact on the model. Otherwise, evidence is either weak or nonexistent to suggest that the other predictors have an effect on predicting the hospital admissions for a day in our dataset. This is another commonality shared with the first model.

In general, while GLMs conceptually seem to be a promising medium for creating predictive models for lockdown prediction and hospital resource demand, they lose their impact in this scenario where interactions are plausible given the situation. The benefit of easy interpretation in lost in some regard. On top of this, we cut many variables to reduce the danger of overfitting.[5] Now we must turn to other methods of modeling to make up for the drawbacks of GLMs, while also hopefully creating a more accurate model as well.

Introducing Machine Learning Models for Predictive Use

We have many concerns when trying to fit a GLM to this data. Thankfully, many of the concerns, like overfitting, can be addressed by building machine learning models using random forest. What is lost in terms of interpretability is regained in terms of accuracy. Using the tidymodels meta-package in R, this process is quite efficient.[6]

When making these machine learning models, it is feasible to fit all of the variables without fear of overfitting and overtraining the model. This also enables us to achieve high levels of accuracy in predicting outcomes in our variables of interest (lockdown status and daily hospital admissions).

The first model, about lockdown status, first novelizes and creates dummy variables for the nominal predictors (month, region). Areas of zero variance are removed, then the numeric variables are all normalized. Lastly, an interaction step is added with all of the six common pollutants to make full use of their potential interactions without worrying about overfitting. Training the model recipe through tidymodels' workflow map system, we are also able to use hyperparameters to their highest function (mtry = 25, min_n = 35) to increase model

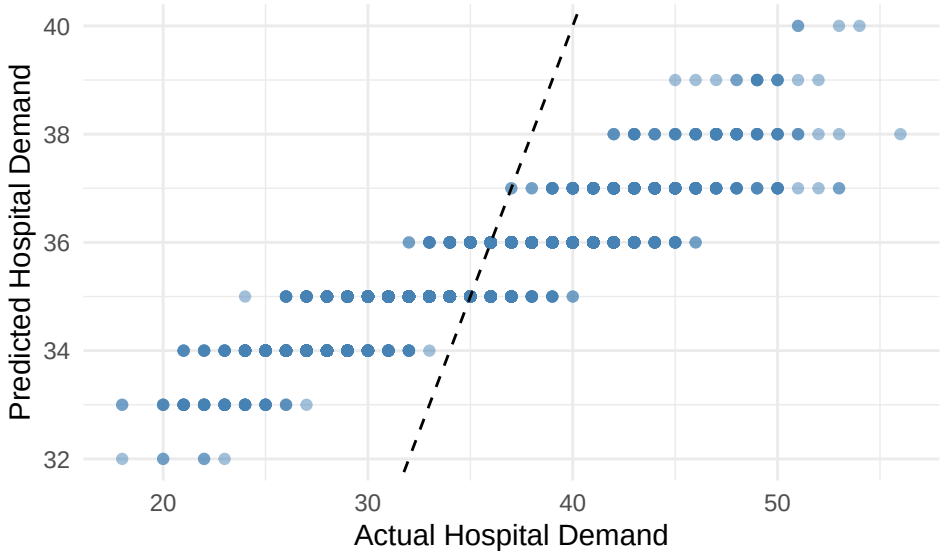
[4] Y = hospital visits + emergency visits + respiratory admissions

[5] Note that both models above only contain three of the six common pollutants. Three were cut for each in process of making both of the above models as they showed no evidence that their coefficients were nonzero.

[6] Version 1.3.0 of the tidymodels meta-package used within R version 4.4.2 "Pile of Leaves".

performance further. What results is a ML model that is able to predict with 85.7% accuracy whether or not there is a lockdown on a random day when testing data is introduced.

The second model focuses on the daily hospital admissions, again drawing that number from multiple variables related to hospital resources. This is another randomforest model which creates dummy variables for the nominal predictors, normalizes numeric predictors, and uses the same interaction terms as the previous model. The hyperparameters mtry and min_n are again in use, and they are set to 1 and 14, respectively. This setup resulted in the lowest rmse across all models (5.92). Though this number is high, it is important to note that this is considering the number of admissions for the entire day. The model can be used to give guidance to hospitals and clinics on how many patients they would come to expect on a singular day. Resources can then be allocated based on its predictions. Since hospitals admit patients throughout the day, rather than all at once, the high rmse is less of a worry, but still worth note.

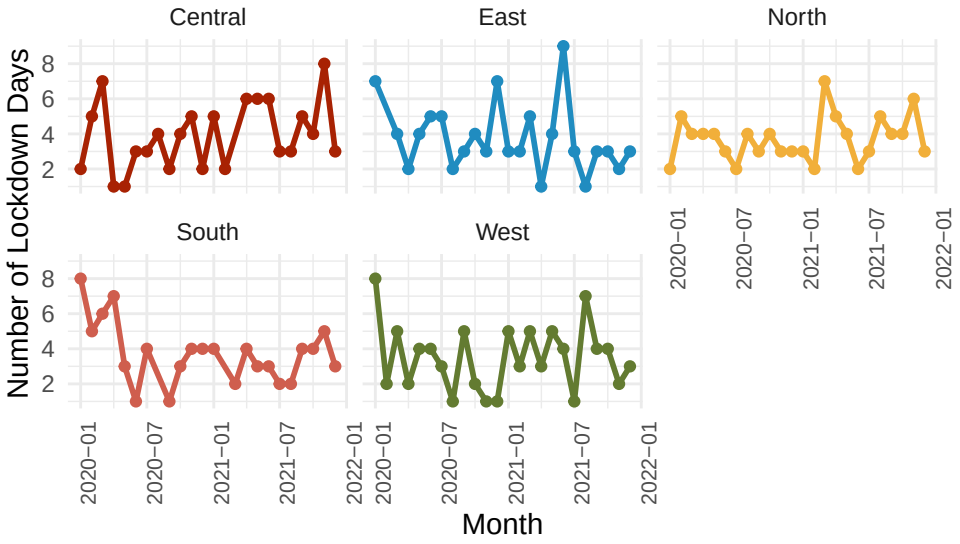


This plot shows the model's consistency in guessing. Note the dashed line, which simply represents the equation $y = x$, or a perfect relationship where the model is 100% accurate (rmse = 0). The consistency at which the model misses the actual hospital demand would indicate bias or improper fit, but I believe this to be the effect of only having one observation per day instead of hourly observations. It shows that there is still room for improvement as well as reason to continue using GLMs in combination with these machine learning mdoels, in the fact that multiple variables have evidence for a noticeable effect on both lockdown status and hospital admissions. Using these machine learning models in conjunction with traditional GLMs where individual variable effect is easier to parse might improve the adaptability and reactive ability of healthcare forces.

Discussion

Creating models for the prediction of lockdown status and hospital admissions based on this synthetic data provided some insight on the explanatory abilities of certain predictor variables while also helping determine what methods actually have the best predictive ability, robustness, and ability to adapt to unseen data. While the original GLMs were easier to interpret on a variable-by-variable basis, machine learning models also prove to be powerful while also being able to consider larger amounts of information. That is not to say that those GLMs are a failure: they emphasize the importance of the six common pollutants as predictor variables, and make it clear that they have interactive potential and work together to impact air quality in such a way that prompts logistical reactions or decreases public health.

There was an interesting lack of explanation to be gathered from regional data. Despite different regional trends, all attempts at modeling showed at best weak evidence for a nonzero effect on the log odds of a lockdown being implemented. Consider the time series below:



Each line represents the number of lockdown days in each given month. We are able isolate region-specific trends because of this, like increasing numbers of lockdown days at the start of the year in only the central and northern regions, as well as overall commonalities, like an increase in lockdowns in the Summer months, likely a cause of wildfires or otherwise increased particle matter levels at that time of year. It is strange to think of the region variable as relatively unhelpful our analysis, but it would seem that that is the case.

There still exists possibilities for the models to have higher predictive power. In the case of hospital admissions, having hourly data might be more helpful and lead to a lower level of error overall. In the current state of the data, the model can only predict admissions on a daily level. This is helpful without a doubt, and would help hospitals and clinics allocate resources, but having hourly predictions could be even more helpful in preparing for waves of demand depending on the time of day-it is highly unlikely that an equal number of patients are admitted to hospitals during every hour of the day.

It is also apparent that machine learning models like these are not infallible at this current moment, at least not in this scenario. There is still reasonable use of traditional GLMs in this scenario. The evidence that they *do* provide on pollutant effect can be helpful in consideration with the information given by predictive models, and should absolutely be used in the future in conjunction with machine learning as long as it has not been perfected.

Data Citation

"Air Quality, Weather, and Respiratory Health", from Khushi Yadav on Kaggle, licensed by MIT.