Allyson Totpal

DSC680

February 15, 2026

Project 2: 10 Questions

1. If ICU occupancy is highly autoregressive, why do we need surveillance signals at all?

Autoregression alone cannot anticipate directional shifts in strain. While lagged ICU levels dominate short-term forecasting, respiratory ED surveillance signals contributed approximately 37-40% of feature importance in nonlinear models. This suggests that surveillance data helps anticipate changes before they fully manifest in occupancy levels, especially at medium-term horizons.

2. Why did linear models fail while Random Forest performed better?

The relationship between ED surveillance signals and ICU occupancy change is likely nonlinear and lag-dependent. Linear regression assumes additive, proportional effects, which may oversimplify the dynamics. Random Forest captures interactions and threshold effects, such as rapid increases in ED visit percentages triggering nonlinear shifts in ICU utilization, leading to improved forecasting performance.

3. Is the model overfitting given the small test set for classification?

Yes, the small test window (n = 14) likely inflates performance metrics such as ROC-AUC. The results should be interpreted as indicative rather than definitive. A rolling or expanding-window cross-validation approach would provide more stable estimates and should be implemented before real-world deployment.

4. Why was a 7-week horizon chosen instead of 3 or 10 weeks?

Three-week forecasts were highly autoregressive and showed minimal added value from surveillance signals. Seven weeks provided a more meaningful operational planning horizon while still maintaining predictive signal. Longer horizons would increase uncertainty and may require additional exogenous variables.

5. Why model statewide data instead of county or hospital-level data.

Statewide aggregation simplifies modeling and reduces noise but also reduces volatility and signal strength. The statewide approach demonstrated proof-of-concept. Future work at county or hospital levels would likely reveal stronger predictive relationships due to increased variation.

6. Could seasonality alone explain the relationship between ED signals and ICU occupancy?

Seasonality likely contributes to the observed relationship. However, the inclusion of lagged Ed signals improves nonlinear model performance beyond autoregressive ICU lags alone, suggesting incremental predictive information beyond seasonal persistence.

7. Why did you evaluate both continuous change and directional classification?

Continuous change forecasting captures magnitude but is sensitive to noise. Directional classification simplifies the problem to operationally relevant decisions. Classification performed better, suggesting that predicting direction is more actionable and statistically stable than predicting exact magnitude.

8. How operationally useful is an 8% RMSE improvement?

In highly persistent systems with low variance, even modest improvements can enhance decision timing. While the absolute RMSE reduction appears small, the directional forecasting results suggest improved identification of rising strain periods, which can meaningfully impact staffing and surge preparation.

9. What happens if a novel respiratory virus emerges?

The model relies on historical relationships between ED signals and ICU occupancy. Structural breaks, like a novel pathogen, could invalidate model assumptions. Ongoing monitoring, recalibration, and integration of additional surveillance sources would be necessary.

10. What are the ethical risks of deploying this model in hospital systems?

Risks include overreliance on probabilistic forecasts, misallocation of resources, and false confidence in model outputs. Transparent uncertainty communication, governance oversight, and integration with clinical judgment are essential to prevent algorithmic decision dominance.