PREDICTING COVID-19 CASE COUNTS PER POPULATION IN THE US (STATE –LEVEL)

DATA MINING, SPRING 2024

CONCORDIA COLLEGE
MOORHEAD, MN
By: Iyanu Lamina

ABSTRACT

The ongoing COVID-19 pandemic has underscored the critical need for predictive models that accurately forecast virus transmission rates and guide public health interventions. In this project, I utilized a diverse set of predictor variables—demographics, mask usage, medical infrastructure, stay-at-home measures, household composition, public health measures, and socioeconomic indicators—to predict COVID-19 case counts per population in various US states.

The aim of this project is to enhance understanding of the determinants driving virus spread and improve prediction accuracy. This project holds significant importance for public health planning, resource allocation, and targeted interventions to mitigate the pandemic's impact. Through comprehensive data analysis spanning pre-COVID-19 and pandemic periods, I plan to uncover patterns and trends to inform proactive responses to emerging COVID-19 dynamics.

OBJECTIVE

The objective of this project is to:

•Develop a predictive model for COVID-19 case counts per population in US states.

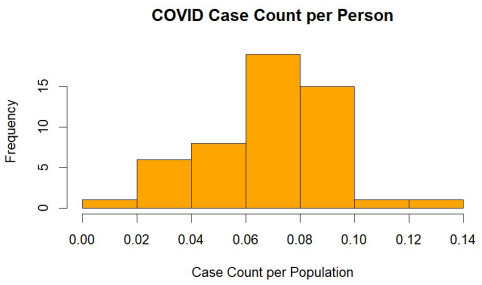
•Integrate diverse predictor variables such as, poverty percentage, annual death from COPD, median household income, total population, and Percentage stayed at home.

•Utilize 2020 data from reputable sources such as the Center for Disease Control (CDC).

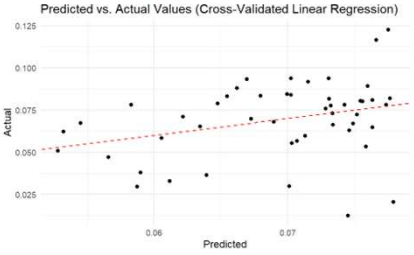
•Analyze pre-COVID-19 and pandemic data to uncover patterns and trends.

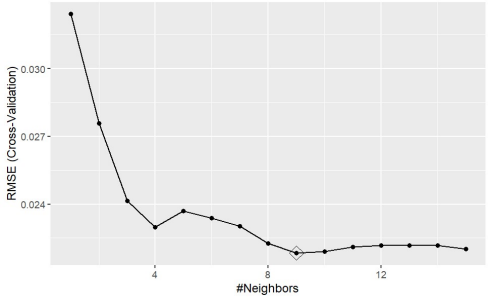
•Inform public health planning and interventions.

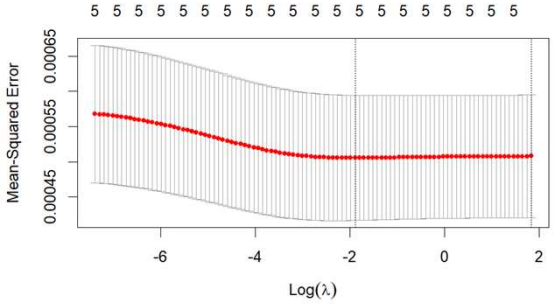
•Aid public health authorities and policymakers in making informed decisions regarding preventive measures, resource allocation, and targeted interventions to control virus spread and mitigate the impact of the pandemic.



GRAPHS







RESULT

Model	RMSE	R-squared	MAE
Linear Reg.	0.02265605	0.07068904	0.01737947
K-NN(k=9)	0.02218933	0.01013269	0.01644634

Ridge	λ	Index	Measure	SE	Nonzero
min	0.151	41	0.0005054	8.883e-05	5
1se	6.239	1	0.0005078	8.720e-05	5

MODEL PERFORMANCE

- Design Decision: Linear regression, ridge regression, and K-Nearest Neighbors each offer unique advantages and can be valuable tools for predicting COVID case count per population, depending on the specific characteristics of the dataset and the underlying relationships between variables.
- From Linear Regression - The model could not perform well due to weak predictors. The results suggest that the current model may not be suitable for making accurate predictions about COVID-19 case counts per population based on the chosen predictor variables. The coefficients of the variables are not statistically significant.
- Ridge Regression -A lambda value of 0.151 suggests moderate regularization, balancing between reducing model complexity and preserving predictive accuracy. The measure for the ridge regression model is 0.0005054.
- KNN -The RMSE for the KNN model is 0.02218933. The best k value is captured at k=9.
- Compare Model performance- Based on the provided metrics, the ridge regression model appears to have superior performance in terms of prediction accuracy compared to the KNN model. Thereby, suggesting better performance in predicting COVID case counts per population for this particular dataset.
- A smaller or lower RMSE or MSE = better performance of a predictive model.

CONCLUSION

In conclusion, this study highlights the significance of integrating diverse predictor variables to accurately forecast COVID-19 case counts in US states. The performance of our predictive models underscores the efficacy of selected variables in capturing transmission complexities. Key factors identified include demographics, healthcare infrastructure, socioeconomic status, and policy measures, emphasizing the importance of timely public health interventions. Geographic location and population density also emerged as significant determinants. Despite limitations, our findings offer valuable insights for evidence-based decision-making and targeted interventions.

Future research should focus on refining models, incorporating additional variables, and conducting longitudinal analyses to monitor transmission dynamics. By advancing our understanding, we can better prepare for future health crises and protect communities nationwide. Additionally, exploring the impact of socioeconomic factors on transmission rates within densely populated states over time is crucial for addressing disparities in outcomes. Likewise, showing a map plot to visualize this.

REFERENCES

Centers for Disease Control and Prevention. (n.d.). *National Environmental Public Health Tracking Network Data explorer*. Centers for Disease Control and Prevention. https://ephtracking.cdc.gov/DataExplorer/

Centers for Disease Control and Prevention. (n.d.-b). *Weekly United States covid-19 cases and deaths by state - archived*. Centers for Disease Control and Prevention. https://data.cdc.gov/Case-Surveillance/Weekly-United-States-COVID-19-Cases-and-Deaths-by-/pwn4-m3yp/about_data