Demographic Predictors of Flu Vaccine Uptake Pre-, During, and Post-COVID-19

Olivia Huang [1] Athena Ke [2] Jacob Hahn [1]

[1]Columbia University [2]Barnard College

Introduction

The COVID-19 pandemic contributed to devastating effects on the healthcare system and led to increased vaccine hesitancy and fatigue. At the same time, vaccination efforts became more politicized, influencing how people view public health recommendations [1]. These changes may have affected seasonal flu vaccination, with uptake patterns decreasing in recent years across different populations. Our project explores how sociodemographic factors relate to these trends.

Research Goal

This project aims to identify the most influential sociodemographic and political factors in determining influenza vaccine uptake rates before, during, and after the COVID-19 pandemic. Existing literature identified critical sociodemographic factors in predicting COVID-19 vaccine uptake [2]. We build on this research by extending the analysis to the flu, by incorporating healthcare and political data, and by following flu vaccine uptake across time to capture broader influences on vaccine behavior.

Data Collection and Processing

We divided our COVID timeline as such: pre-COVID (2016-2020), during COVID (2020-2022), and post-COVID (2022-2025). For each time period, we compiled a data set by aggregating county-level data from the U.S. Census Bureau and annual County Health Reports using FIPS codes. Alaska, Connecticut, and Louisiana were excluded due to data unavailability.

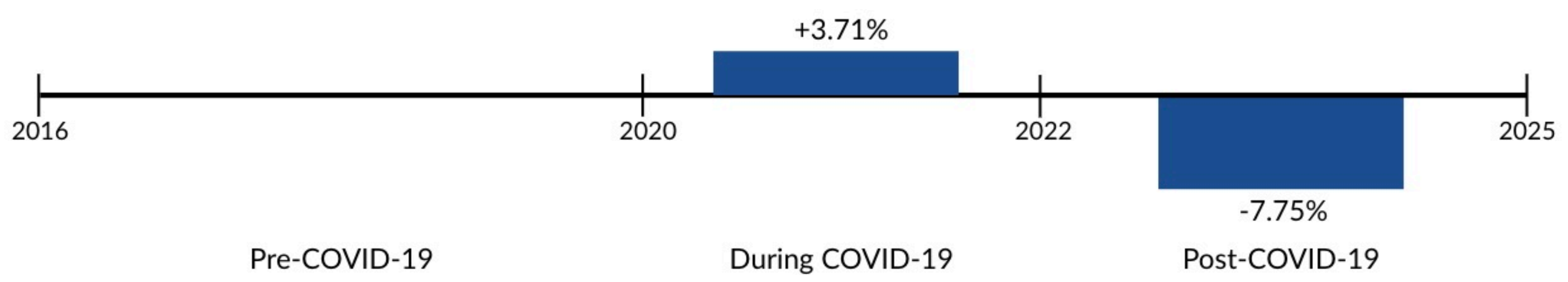

Figure 1. Timeline used to orient data collection with percent change in the national influenza vaccination rate compared to the previous time period.

We identified 45 demographic features to act as predictors of vaccine uptake. The features may be divided into general categories: *Healthcare Access, Political Representation, Socioeconomic Status and Poverty, Education, Occupation, Housing Status, Age, Race*, and *Disability Status*.

Methodology

Ridge Regression and LASSO

Ridge regression addresses multicollinearity in ordinary least squares regression by introducing a ridge parameter [3], while LASSO adds a penalty to the residual sum of squares, shrinking weakly correlated coefficients to zero [4]. We applied both methods across each time period, using their coefficients for comparison and their prediction errors as a baseline for evaluating a more comprehensive model.

XGBoost

eXtreme Gradient Boosting (XGBoost) is a supervised machine learning algorithm that applies gradient-boosted decision trees to regression and classification problems while seeking to minimize a loss function [5]. We used XGBoost to create one regression model for each time period. We repeatedly imposed a random 80-20 train-test split and took the set of hyperparameters that produced the lowest RMSE. Using this set as the model pipeline, we retrained the model on the entire dataset and calculated performance metrics.

We visualized the model fitted to the entire dataset using choropleth maps, with unavailable regions grayed out.

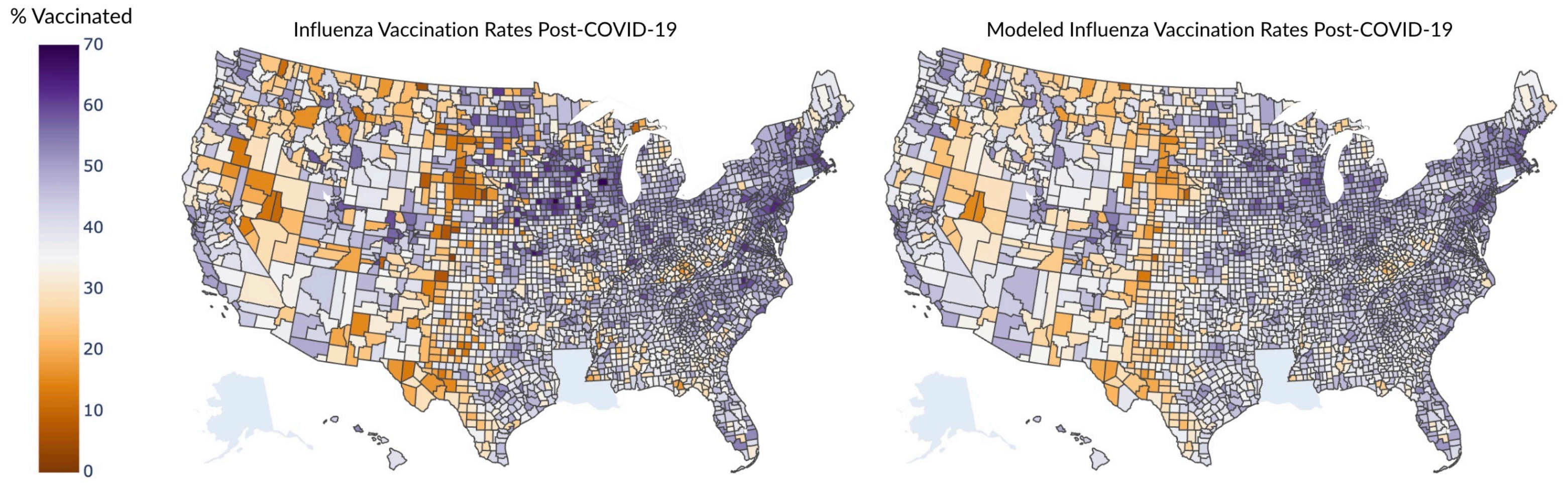
% Vaccinated
70
60
50
40
30
20
10
0
Influenza Vaccination Rates Post-COVID-19
Modeled Influenza Vaccination Rates Post-COVID-19

Figure 2. Actual vs. modeled vaccination rate generated by XGBoost for post-COVID period with an RMSE of 4.07.

Error and Feature Importance

We used adjusted R-squared, RMSE, and MAPE to evaluate performance. Feature importance was determined using SHapley Additive exPlanations (SHAP). Specifically, TreeSHAP, outlined in equation (1), is a game-theoretic approach that explains how a change in an outcome with respect to a baseline may be attributed to different input features in models of tree structure [6].

$$\phi_i(x) = \sum_{S \subset F\{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right] \tag{1}$$

Where: x is the observation input, $\phi_i(x)$ is the Shapley value for feature i for input x under the model f, F is the set of all features, f_S is the trained model on the subset of features S, $f_{S\cup\{i\}}$ is the trained model on the subset $S\cup\{i\}$, x_S is the restricted input of x given subset S, and $x_{S\cup\{i\}}$ is the restricted input of x given subset $S\cup\{i\}$.

Results

The post-COVID model performed the best, with an RMSE of 4.07, MAPE of 9.01, and adjusted R-squared of 0.84. To assess category-level influence, we used both a mean reciprocal rank and a sum of SHAP values to aggregate specific feature contributions.

- **Pre-COVID**: Political landscape, race, and education were the strongest predictors of flu vaccine uptake.
- **During COVID**: Occupation gained predictive power; race and education remained significant.
- **Post-COVID**: Socioeconomic status became significant, political landscape resurfaced, and occupation remained.

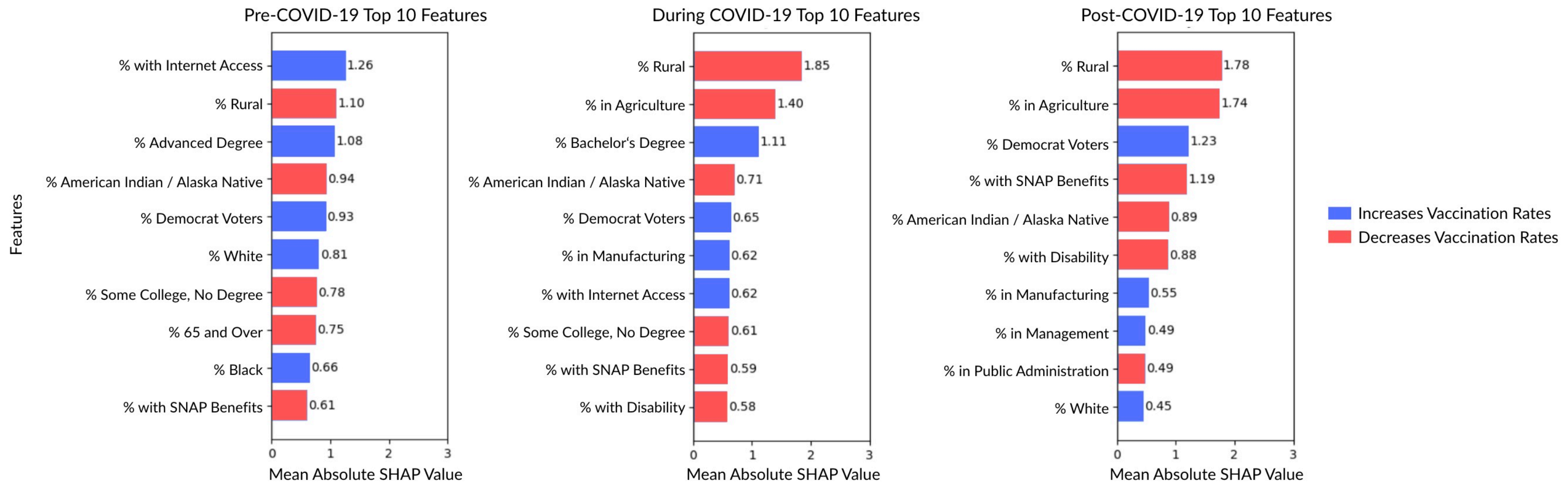

Figure 3. Mean absolute SHAP values with added directionality for pre-, during, and post-COVID general models

Granular Analyses

"Percent Rural" (percentage of county population living in a census-defined rural area) was consistently a top predictor, leading us to stratify the data by urban, mixed, and rural counties. For each time period, we created separate models for each group.

- In **urban** areas, the top predictors changed over time. Only socioeconomic status and race remained consistent; occupation gained significance.
- In counties with an **urban-rural mix**, race and political landscape were consistently significant. Occupation emerged as a top predictor in later periods.
- In **rural** areas, five of the top ten predictors remained stable across time, primarily relating to race, occupation, socioeconomic status, and political landscape.

Table 1. Model performance metrics by area and time period.

Area	Time Period	RMSE (%)	MAPE (%)	Adjusted R^2
Urban	Pre-COVID	1.70	2.75	**0.94**
	During COVID	1.09	1.75	**0.98**
	Post-COVID	0.63	1.02	**0.99**
Urban-Rural Mix	Pre-COVID	4.14	8.31	**0.78**
	During COVID	3.41	6.12	**0.84**
	Post-COVID	3.67	7.06	**0.81**
Rural	Pre-COVID	6.73	18.50	**0.56**
	During COVID	6.16	15.84	**0.66**
	Post-COVID	5.36	15.30	**0.71**

Conclusion and Implications

Our granular XGBoost models were less generalizable to filtered data, while the three initial models may be more broadly applicable to different sets of data. SHAP analyses revealed several key patterns:

1. 'Traditional' structural and socioeconomic barriers to vaccine access (such as education, Internet access, and race) were less significantly linked to vaccine uptake rates.
2. Local political landscapes and occupation are increasingly correlated to vaccination behavior.

The shift in predictors of vaccine uptake suggests that COVID-19 transformed vaccination from an issue of access into one of personal belief and identity. To address this pressing issue, public health officials must move beyond access-focused strategies and instead engage belief systems, rebuild trust, and implement culturally resonant, community-driven approaches.

References

[1] Su Z, Cheshmehzangi A, McDonnell D, da Veiga CP, Xiang YT. Mind the "Vaccine Fatigue". Front Immunol. 2022 Mar 10;13:839433. doi:10.3389/fimmu.2022.839433.

[2] Cheong Q, Au-Yeung M, Quon S, Concepcion K, Kong JD. Predictive Modeling of Vaccination Uptake in US Counties: A Machine Learning-Based Approach. J Med Internet Res. 2021 Nov 25;23(11):e33231. doi:10.2196/33231.

[3] Columbia Mailman School of Public Health. Ridge Regression. https://www.publichealth.columbia.edu/research/population-health-methods/ridge-regression. Accessed July 8, 2025.

[4] Columbia Mailman School of Public Health. Least Absolute Shrinkage and Selection Operator (LASSO). https://www.publichealth.columbia.edu/research/population-health-methods/least-absolute-shrinkage-and-selection-operator-lasso. Accessed July 8, 2025.

[5] Wiens M, Verone-Boyle A, Henscheid N, Podichetty JT, Burton J. A Tutorial and Use Case Example of the eXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications. Clin Transl Sci. 2025;18(3):e70172. doi:10.1111/cts.70172.

[6] Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2:56-67. doi:10.1038/s42256-019-0138-9.

[7] Fadel S. Explainable machine learning, game theory, and Shapley values: A technical review. Statistics Canada. February 28, 2022. https://www.statcan.gc.ca/en/data-science/network/explainable-learning.

[8] CDC. Influenza vaccination coverage for persons 6 months and older. May 18, 2021. https://www.cdc.gov/fluvaxview/interactive/general-population-coverage.html.

oh2251@columbia.edu, yk3110@barnard.edu, jrh2252@columbia.edu

Columbia Summer Undergraduate Experience in Math Modeling 2025

Advised by George Dragomir, Vihan Pandey, Dobrin Marchev