Investigating the Influences and Dynamics:

A Trend Perspective on COVID-19 Booster Uptake in the United States

Abstract

This project focused on the trends in COVID-19 vaccine booster uptake in the United States. It replicated and built upon the model from the paper "Modeling for COVID-19 College Reopening Decisions: Cornell, a Case Study," which utilized epidemiological models to predict pandemic progression, aiding in safer reopening decisions, as exemplified by Cornell University. The study "Investigating the Influences and Dynamics: A Trend Perspective on COVID-19 Booster Uptake in the United States" delved into the impact of various parameters on booster vaccine doses, examining the relationship between long-term infection prevention and people's willingness to receive cumulative vaccine doses from 2023 to 2026. It employed logistic regression and ARIMA models to analyze statistical significance and explored trends related to age, race, and geography in vaccine uptake, as well as people's intentions. The model's accuracy was approximately 0.8, effectively reflecting the variety in trends.

2.Model Logistic Regression

The Logistic Regression model is a statistical method employed for predicting the probability of an event occurrence based on one or more predictor variables. The mathematical expression of the Logistic Regression model is as follows:
Let Y be the binary outcome variable (0 or 1), and $X_1, X_2, \ldots, X_n$ be the predictor variables. The model assumes a linear relationship between the predictor variables and the log-odds of the event:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Here, p represents the probability of the event occurring, and $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients associated with each predictor variable. The logistic function is then applied to transform the linear combination into the probability scale:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

3.Model ARIMA

An Autoregressive Integrated Moving Average (ARIMA) Prediction Model applied statistical analysis using an integrated moving average. It predicted future trends based on past data performance.
A process $\{X_t\}$ followed an Integrated ARIMA model, denoted by ARIMA(p, d, q).

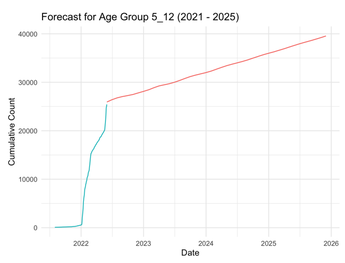
$$\text{If } \nabla^d X_t = (1-B)^d X_t \text{ is ARIMA}(p,q), \text{ we write the model as}$$
$$\phi(B)(1-B)^d X_t = \theta(B) Z_t,$$
$$Z_t \sim WN(0, \sigma^2).$$

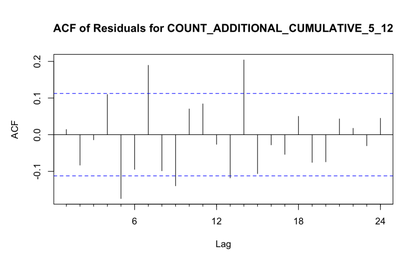
The integration parameter d is a non-negative integer.
The ARIMA model utilized the `auto.arima()` function in R, with cross-validation performed using `tsCV()`. The prediction forecast one step ahead using $h = 1$, assuming 12 observations per time unit. The forecasts were stored in a list, indexed by age groups. Error metrics computed for each cross-validation fold were also stored in a list corresponding to each forecast. The results were interpreted by two lines representing historical and forecast data.

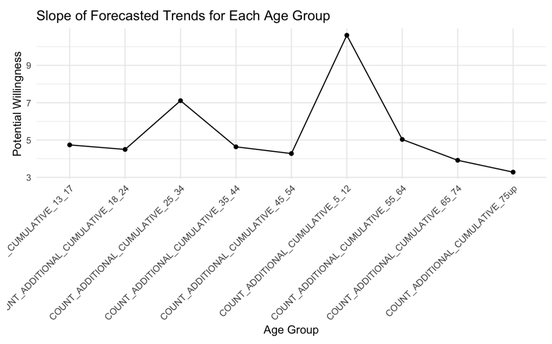
5:Results

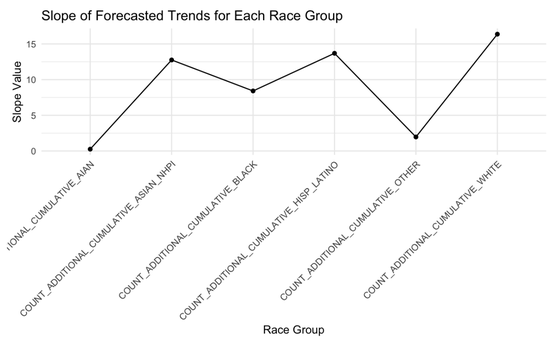
- The Logistic Regression results suggest a higher likelihood of future COVID booster acceptance among the **White** population, followed by **Hispanics**. Train Accuracy: **81.61%** Test Accuracy: **82.29%** Pseudo-R square : **0.8475**
- The ARIMA model highlights a potentially increased tendency for age groups to receive COVID booster shots between **2023** and **2026**. Result: **5-12** and **25-34** age groups are significant.
- The ACF in the ARIMA forecast showed an exponentially decaying pattern. It analyzed correlations up to **24 time** periods back, with autocorrelation coefficients between **-0.1 and 0.2**, indicating the linear relationship strength and direction. The blue dashed line marked significance. Mostly, bars stayed within this line, implying little significant autocorrelation. The residuals, resembling **white noise**, suggested the model effectively captured the data.

race <chr>	mean_probability <dbl>
Hispanic	0.4483241
Multiple Races	0.3035726
NH American Indian or Alaska Native	0.3684169
NH Asian or Pacific Islander	0.3819873
NH Black	0.3692817
White	0.4911026





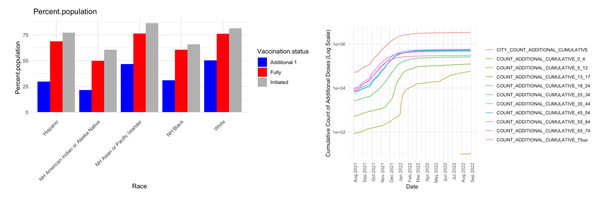




1.Data Description:SARS-CoV-2 Vaccine

COVID-19 Updated (Bivalent) Vaccination Coverage By Race/Ethnicity and Age Group

- This table presents the cumulative number and percentage of individuals aged 5 years and older in CT and NYC who have received an updated (bivalent) COVID-19 vaccination, categorized by race/ethnicity and age group. The race and ethnicity data were self-reported and sourced from existing electronic health care records. Last Updated: August 7, 2023. Table Dimensions: 1680 Rows x 13 Columns.

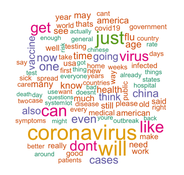


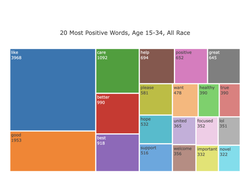
COVID-19 vaccinations given by NYC facilities and reported to the Citywide Immunization Registry (CIR)

- This table categorizes age groups according to U.S. Census age categories and includes demographic information, such as race/ethnicity, sex, and gender identity, related to recipients of the Pfizer-BioNTech, Moderna, and Johnson Johnson vaccines. Data on second booster doses is included as of June 10, 2022.

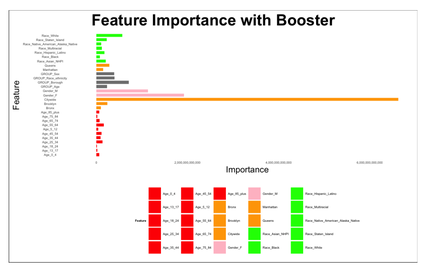
4.Error Analysis

- **Sentiment analysis(nlp)** was introduced to examine the qualitative aspects of individuals' attitudes towards COVID-19 vaccination, based on **Reddit comments** from 2022 to 2023, further exploring the emotional tone.
- The **increase in cumulative counts** suggested an objective approach to vaccination but **did not directly correlate with positive intentions**.
- The **sentiment scores, overall positive**, aligned with logistic and ARIMA model predictions for specific age groups and ethnicities, reinforcing the consistency of findings across different analytical approaches.





- **Random forest model** analysis identified **geographical factors** as key in predicting COVID-19 booster acceptance, stressing the need for diverse parameters for comprehensive results.



6. Acknowledgements

- **Special thanks to Professor Donna Slonim and Teaching Assistant Hao Zhu**. For any questions, please reach out to Jiatai Zhang (jzhang52@tufts.com) and Jiayi Zhang(jzhang47@tufts.edu).
- Data, code, and references are available at https://github.com/zjh0042/CS169-Final-Project.git.