Predicting COVID-19 Vaccine Hesitancy

William Cheng

December 19, 2021

Contents

The code to reproduce this report is available on Github.

1 Executive Summary

Problem. Despite the growing body of research focused on understanding COVID-19 vaccine hesitancy, there is still much to be learned about the factors that make some individuals more susceptible to COVID-19 vaccine hesitancy. Hence, for my final project, I decided to look into a wide range of social, economic, health, and lifestyle factors and analyze which factors were most predictive of vaccine hesitancy in individuals. In this project, I analyzed data from the most recent collection period (September 29, 2021 - October 11, 2021) of the Household Pulse Survey conducted by the U.S. Census. While only analyzing data from this period inherently limits my results, utilizing data within this short timeframe generally controls for a variety of external events that can affect vaccine hesitancy such as federal/employer vaccine mandates or news about the spread of new COVID variants.

Data. My dataset pulled data from the Household Pulse Survey, a program by the U.S. Census. My explanatory variables span four main categories of factors: social (e.g. age, ethnicity, marital status), health (e.g. anxiety, depression), economic factors (e.g. health insurance coverage, income), and lifestyle factors (e.g. working from home, eating at restaurants). My primary response variable of interest was the hesitancy to get the COVID-19 vaccine.

Analysis. Before exploring the data or running any analyses, I split the data into a training dataset and a test dataset, with the training dataset used for building the predictive models and the test dataset used for assessing and comparing model performance. Then, I explored the data to check for class imbalance and to assess correlations between variables and between variables and the response. In order to build an optimal predictive model, I constructed six different models: logistic regression, ridge regression, LASSO regression, elastic net regression, classification tree, and boosting. Of the regression models, logistic regression had the lowest misclassification rate and of the tree-based models, the boosted model had the lowest misclassification rate. The logistic regression and the boosted model had near-identical misclassification rates.

Conclusions. Interestingly, I found that the logistic regression and boosted model both pointed to similar types of variables as the strongest predictors of vaccine hesitancy. Specifically, the boosted model revealed that age, education level, and prior exposure to COVID emerged as the most significant predictors. As herd immunity is only achieved when the majority is vaccinated, public health officials can use these findings to adjust their messaging to better encourage individuals in specific populations to receive the COVID vaccines. I hope that this analysis can inform additional policies aimed at improving vaccination rates in the United States, both in the context of the COVID-19 pandemic and more generally going forward.

2 Introduction

Background. The ongoing pandemic of coronavirus (SARS-CoV-2) is already one of the deadliest pandemics in recent history. Since its initial discovery in 2019, the virus has infected over 274,628,461 million people and killed over 5,358,978 worldwide [1]. In the United States, the virus has infected 51,110,283 people and killed 806,335 as of December 21, 2021 [2]. The pandemic has also led to one of the fastest developments of a vaccine as well as one of the largest vaccine rollouts in history [3]. That being said, there has been an increased spread in misinformation related to the COVID-19 vaccines and there are certain populations in the United States who have resisted getting vaccinated. The low vaccination rates among these populations can have negative impacts on herd immunity and can lead to the increased spread of coronavirus variants[4]. A thorough analysis of individuals who are hesitant about getting vaccinated can help inform strategies to improve vaccination rates within the United States to mitigate the COVID-19 pandemic as well as future disease outbreaks.

Past research has shown that vaccine hesitancy is a complex issue and can be influenced by a variety of factors. For instance, the World Health Organization outlined several determinants of vaccine hesitancy, including media environment, influential leaders, historical influences, geographic barriers, and the socio-economic

[1] https://covid19.who.int/
[2] https://www.nytimes.com/interactive/2021/us/covid-cases.html
[3] https://connect.uclahealth.org/2020/12/10/the-fastest-vaccine-in-history/
[4] https://www.cbsnews.com/news/fauci-covid-omicron-variant-herd-immunity-vaccine-infection/

factors [5]. Despite these efforts to outline a framework for vaccine hesitancy, there is still much to be learned about the drivers of vaccine hesitancy because many of these determinants can be difficult to measure.

Analysis goals. Given the preexisting knowledge of the capacity for a variety of factors to influence vaccine hesitancy, I sought to investigate how social, economic, health, and lifestyle factors in particular influence vaccine hesitancy. Specifically, I was interested in which kinds of factors (social, economic, health, and lifestyle) and which specific variables were most predictive of vaccine hesitancy in individuals.

Significance. I hope that my analysis will contribute to the growing body of research on factors contributing to COVID-19 vaccine hesitancy, and I hope that they can be used to improve the efficacy of future vaccination campaigns. My results highlight the importance of analyzing social, economic, health, and lifestyle factors in efforts to understand vaccine hesitancy.

3 Data

3.1 Data sources

My dataset originated from the Household Pulse Survey conducted by the U.S. Census[6]. The Household Pulse Survey seeks to measure how the COVID-19 pandemic is impacting households across the country from a social and economic perspective. Specifically, along with measuring the intention of the respondent to get vaccinated for COVID-19, the Household Pulse Survey includes measures of demographics, economic welfare, and lifestyle factors. The Household Pulse Survey was conducted through phone calls and online surveys that asked the respondent a set list of questions relating to the above measures. To construct my dataset, I pulled data from one collection period, amounting to the timeframe between September 29 to October 11, 2021.

3.2 Data cleaning

My first task in the data cleaning phase of the project was clean up the features in the raw dataset. As the Household Pulse Survey was conducted as a survey, the levels of some variables were spread out through multiple columns of the data, creating redundancy issues. I reorganized several variables and I reduced the levels of several factor features. The second task in the data cleaning phase was the recoding of my response variable, which is further explained in the Response Variable section below. Next, I realized that there was a significant class imbalance in the raw data. The ratio of observations who were not hesitant to observations who were hesitant was 11 to 1. Thus, I utilized down-sampling so the ratio then became a more manageable 1.65 to 1.

3.3 Data description

3.3.1 Observations

My cleaned dataset contained a total of 10,230 observations, each corresponding to an individual respondent.

3.3.2 Response Variable

My binary response variable signifies if an individual is hesitant or not hesitant about getting vaccinated for COVID-19. I created this variable from two key measurements related to vaccine hesitancy that the Household Pulse Survey collected. The first measured if the respondent had already taken the COVID-19 vaccine, and if not, the second measured if the respondent intended on taking the vaccine. The second measurement had five separate levels: definitely, probably, unsure, probably not, or definitely not getting vaccinated. Thus, my recoded response variable was categorical: an individual could be Hesitant (1) or Not Hesitant (0) about getting vaccinated for COVID-19. An observation was considered hesitant if the respondent had not taken the vaccine and signified that they were unsure about getting vaccinated, probably not going to get vaccinated, or definitely not going to get vaccinated. An observation was considered not

[5] https://www.who.int/immunization/sage/meetings/2013/april/1_Model_analyze_driversofvaccineConfidence_22_March.pdf

[6] https://www.census.gov/programs-surveys/household-pulse-survey/datasets.html

hesitant if the respondent had received at least one dose of the vaccine or signified that they were probably or definitely going to get vaccinated.

3.3.3 Features

Drawing on the data from the Household Pulse Survey, I included 62 explanatory variables in my analysis, which fall into four main categories: social, economic, health, and lifestyle. For a detailed specification of these variables, refer to Appendix A.

3.4 Data allocation

I split my dataset into two subsets: a training dataset used for building the predictive models and a test dataset used for evaluating the models. I utilized an 80-20 split, such that the training dataset consists of 80% of the observations and the test dataset consists of 20% of the observations. I set a random seed to ensure that any splits done would lead to the same results.

3.5 Data exploration

3.5.1 Response

I first sought to understand the distribution of the response variable in the training dataset. As seen in Figure 1, after downsampling to adjust for class imbalance, there are 3082 observations that are vaccine hesitant, comprising 37.7% of the data, while there are 5102 observations that are not hesitant, comprising 62.3% of the data.

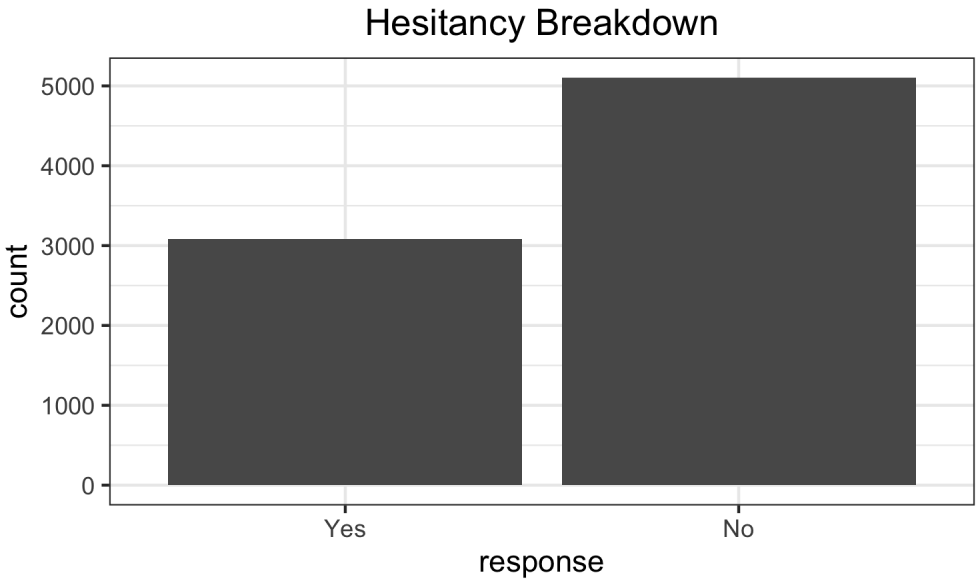

Figure 1: Distribution of responses

3.5.2 Features

Three of my features were numerical (age, number of kids in household, and number of adults in household), and so I sought to look the relationship between these features. A correlation plot can be seen in Figure 2, and it seems to suggest that there a slight negative correlation between age and the number of children in a household.

To explore the relationship between these three numerical variables and the binary response, I constructed three boxplots, which can be seen in Figure 3. In the first plot between vaccine hesitancy and age, the median

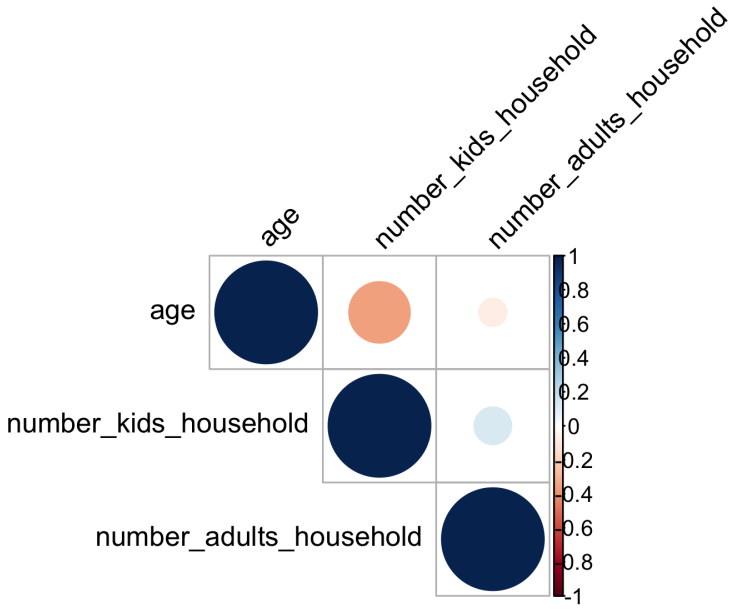

Figure 2: Relationship between numerical features

age of those who are hesitant (46 years) is lower than the median age of those who are not hesitant (58 years). In the second plot between hesitancy and the number of children in the household, although the median number of children of both hesitant and non-hesitant individuals was both 0, there was a larger interquartile range for individuals who are hesitant. The same analysis applied to the third boxplot between hesitancy and the number of adults in the household, as the median number of adults in the household was around 2 for both hesitant and not hesitant populations, there was a larger interquartile range for individuals who are hesitant.

As my response was categorical and all my features apart from three (age, number of kids in household, and number of adults in household) were categorical with more than two levels, I decided to construct a Cramer's V association matrix for all non-numerical factors, grouped by the different categories of factors: social, economic, health, and lifestyle.

The matrix for social factors can be seen in Figure 4. The matrix shows that age and education level have the largest positive associations with vaccine hesitancy. There are positive associations between age and marital status, described gender and birth gender, and sexual orientation and described gender.

The matrix for economic factors can be seen in Figure 5. The matrix shows that the variables of having kids, having difficulty with expenses, household food sufficiency, having trouble with energy bills, and having children who attend school have the largest positive associations with vaccine hesitancy. There are strong associations between several variables related to children, including having children and receiving child tax credit, and between having children and sending children to public, private, or homeschool. There are also associations related to educational variables, such as between having children attending school and having at least one member of the household participate in summer educational catch-up activities. Finally, there are associations related to housing variables, namely between housing type and being caught up on rent or mortgage payments.

The matrix for health factors can be seen in Figure 6. The matrix shows that the variables of having prior exposure to COVID and having a child who missed health checkups have the largest positive associations with vaccine hesitancy. There are strong associations between several variables related to mental health here, namely between feeling anxious, feeling worries, losing interest, feeling depressed, being on prescription

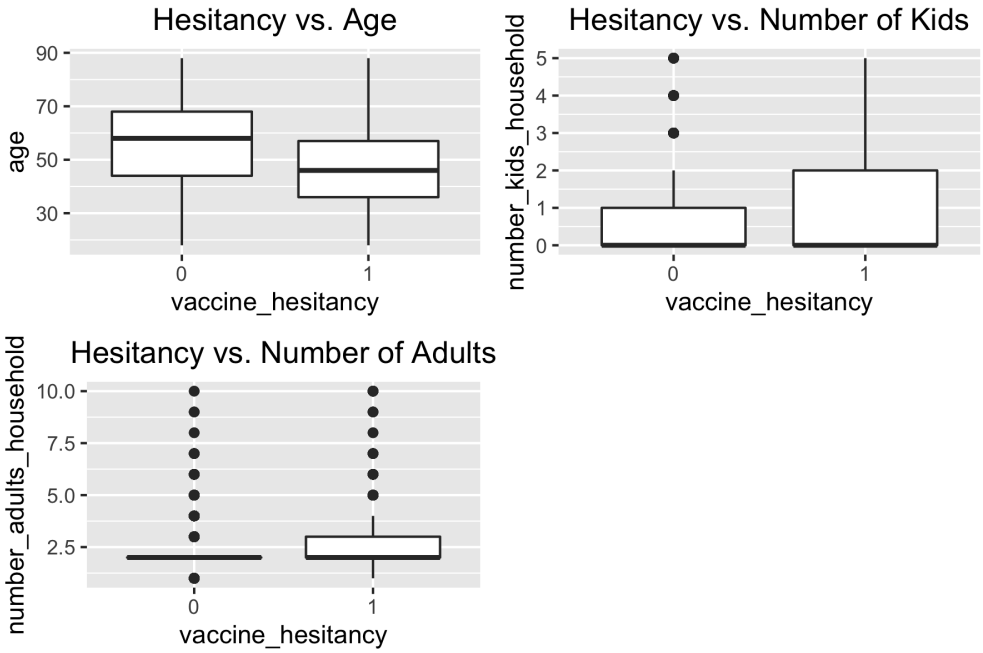

Figure 3: Relationship between numerical features

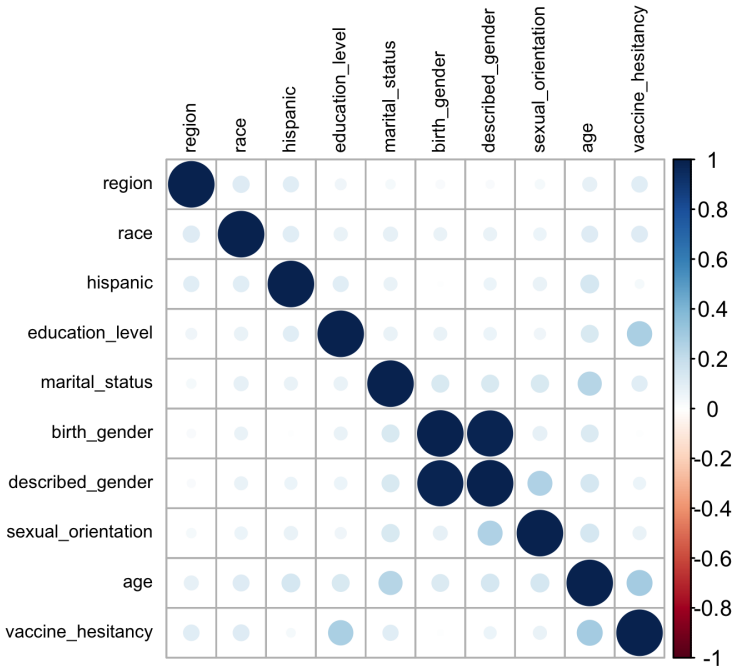

Figure 4: Association between different social variables and the response

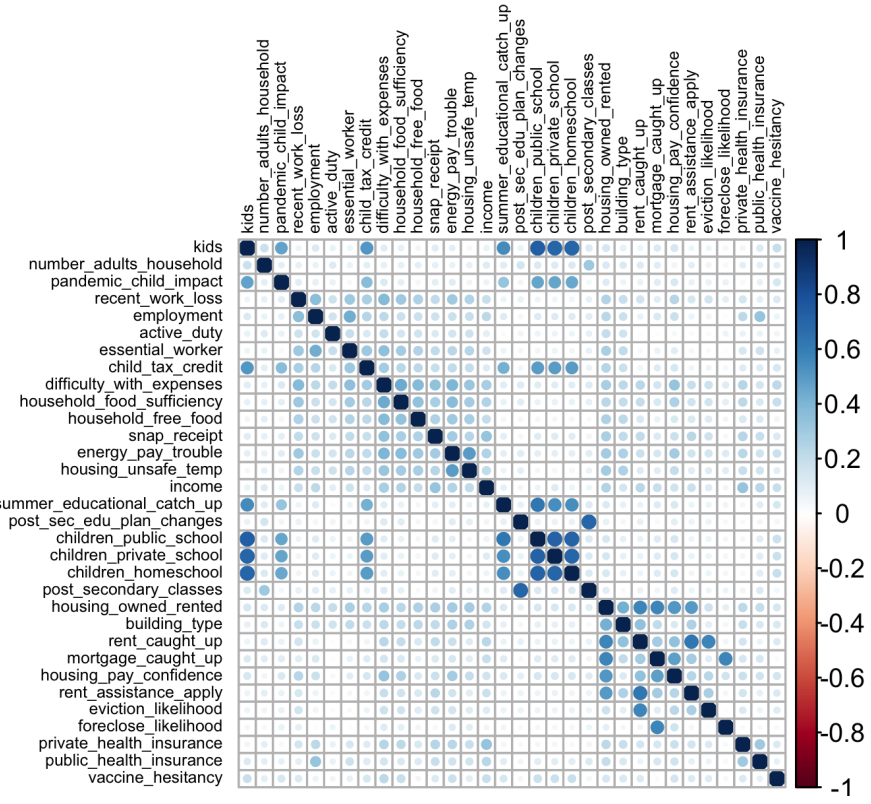

Figure 5: Association between different economic variables and the response

mental health treatment, receiving mental services, or needing mental services and not receiving them. There are also strong associations between the variables regarding physical health, including seeing limitations, hearing limitations, remembering limitations, and mobility limitations.

The matrix for lifestyle factors can be seen in Figure 7. The matrix shows that the variables of working and having a child who uses telehealth have the largest positive associations with vaccine hesitancy, although they are not very strong. There are strong associations between several variables here, namely between working onsite, working from home, in store shopping, eating indoors in restaurants, having housekeeping, and having in person medical appointments.

4 Modeling

4.1 Regression-based methods

4.1.1 Logistic regression

For the first model, I decided to fit a logistic regression with all 62 features to the training data. The logistic regression revealed that the following variables are significantly associated with the response at the 0.05 level: living in specific regions (South, Midwest, and West), being of Hispanic origin, being Black or Asian, having a bachelor's or graduate degree, sexuality (LGBTQ+), having a high number of people in the household, having prior exposure to COVID or potential prior exposure, having no to little difficulty with expenses, shopping in stores, receiving and using a SNAP receipt, having no private or public health insurance, living in a mobile-style home (mobile, boat, RV, van, etc.), having higher income, having to personally take on additional care for children, having at least one child attend private school or homeschool, and age.

For the logistic model, Table 1 shows the resulting confusion matrix. The performance metrics are listed in Table 2, shows that the model has a misclassification error rate of 0.25. The resulting false positive rate is

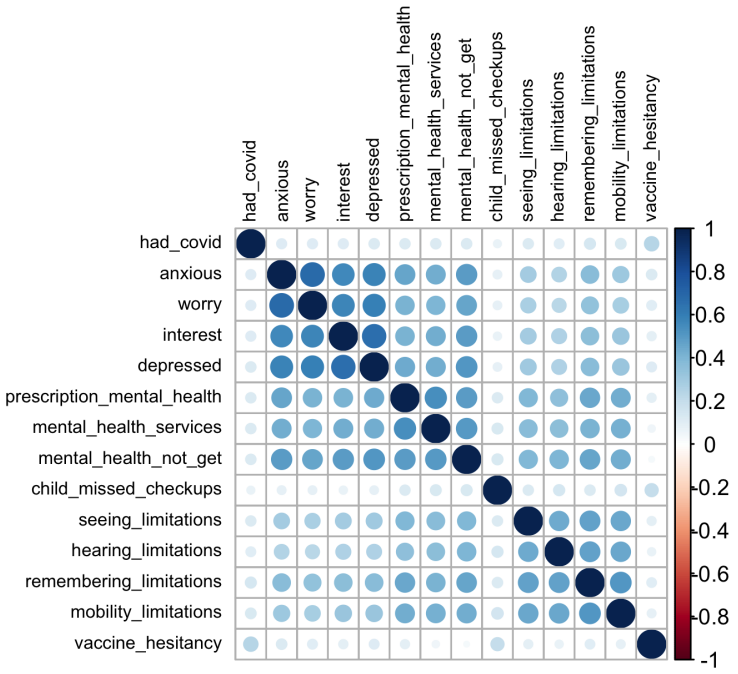

Figure 6: Association between different health variables and the response

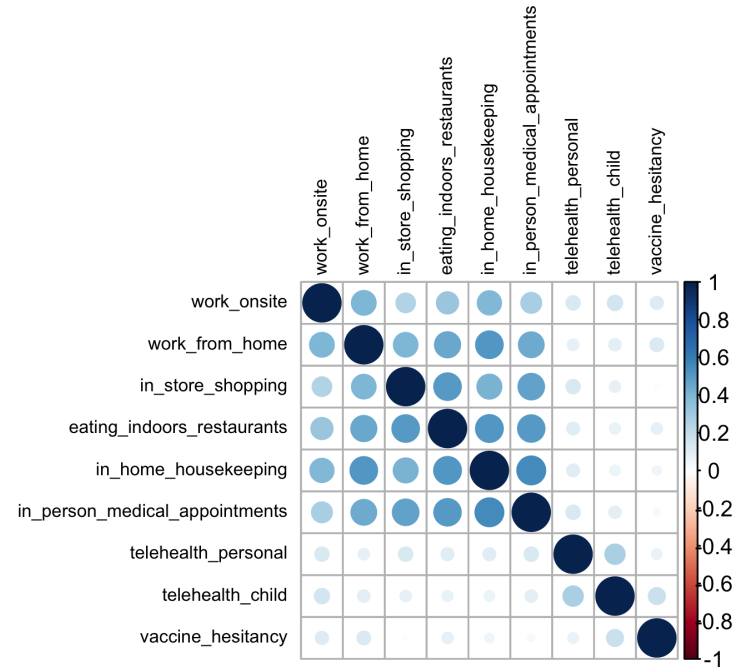

Figure 7: Association between different lifestyle variables and the response

0.17 and the false negative rate is 0.39, which shows that the model could have lower sensitivity and higher specificity.

Table 1: The confusion matrix for the logistic regression

	0	1
0	1058	211
1	300	477

Table 2: The performance metrics of the logistic regression

Metric	Classifier Performance
Misclassification Error	0.25
False Positive Rate	0.17
False Negative Rate	0.39

4.1.2 Ridge regression

While the logistic regression seemed to produce a model that worked well, fitting a logistic model with many explanatory variables can lead to high variance and suboptimal predictions. Thus, I decided to build and evaluate shrinkage models with the aim of reducing variance and obtaining a more accurate model. I ran three cross-validated regressions for which optimal values of lambda were chosen according to the one-standard-error rule: ridge regression, LASSO (Least Absolute Shrinkage and Selection Operator) regression, and elastic net regression.

For the ridge regression, Figure 8 shows the CV plot, Figure 9 shows the trace plot, and Table 3 lists the top 10 features with the highest standardized coefficients. As seen from the trace plot and table of coefficients, the variables of age, education level, and having prior exposure to COVID seemed to have the greatest association with the response. Specifically, the older an individual is, the less likely they are to be vaccine hesitant. Those with higher educational attainment and those who did not have prior exposure to COVID were also less likely to be vaccine hesitant. Individuals with lower educational attainment as well as those with prior exposure to COVID are more likely to be vaccine hesitant.

The confusion matrix can be seen in Table 4 and the performance metrics for the ridge regression can be seen in Table 5. The ridge model has a misclassification rate of 0.27, which is higher than the logistic regression. According to the false positive rate of 0.14 and the false negative rate of 0.48, the ridge regression seems to have higher specificity and lower sensitivity than the logistic regression.

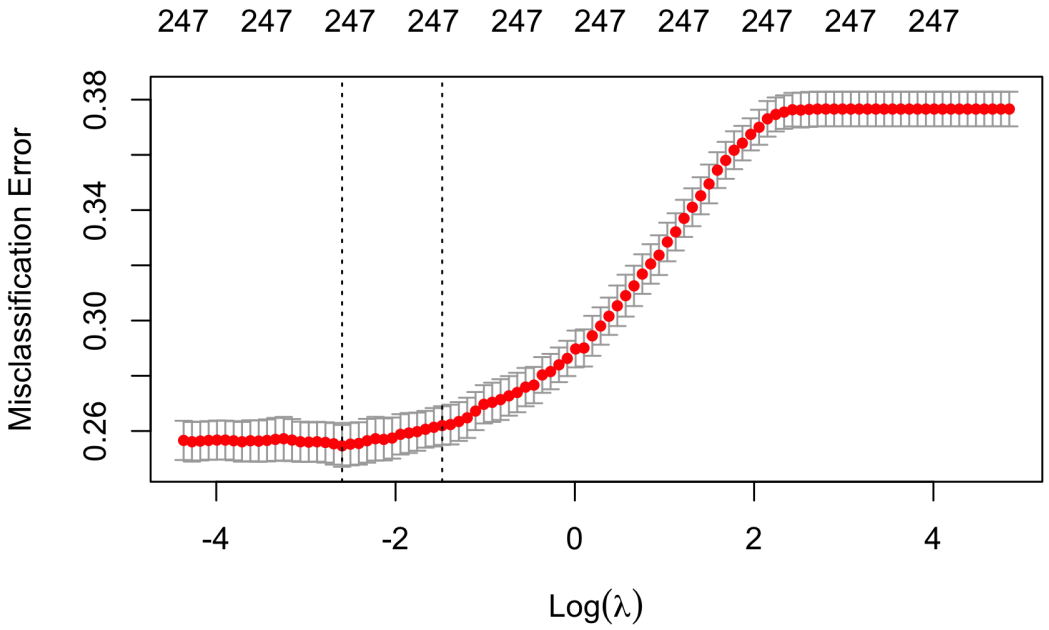

Figure 8: Ridge CV plot.

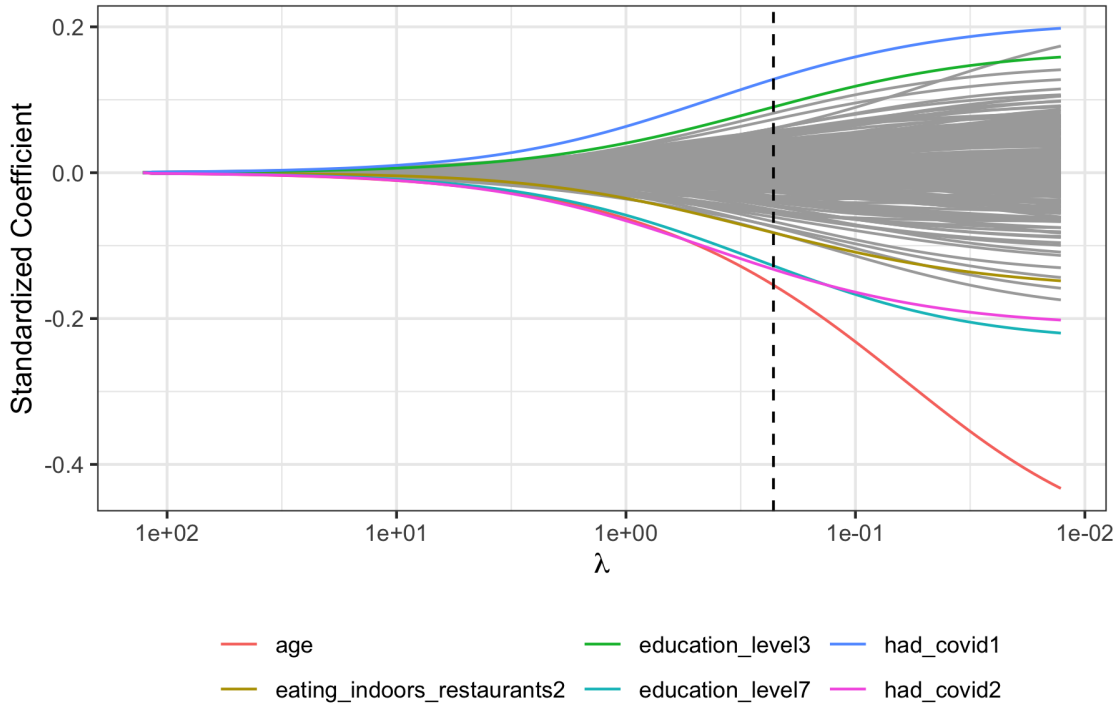

Figure 9: Ridge trace plot.

Table 3: Standardized coefficients for the top 10 features in the ridge model based on the one-standard-error rule.

Feature	Coefficient
age	-0.15
had_covid2	-0.13
had_covid1	0.13
education_level7	-0.13
education_level3	0.09
eating_indoors_restaurants2	-0.08
income8	-0.08
eating_indoors_restaurants1	0.08
race3	-0.07
difficulty_with_expenses1	-0.07

Table 4: The confusion matrix for the ridge regression

	0	1
0	1097	172
1	373	404

Table 5: The performance metrics of the ridge regression

Metric	Classifier Performance
Misclassification Error	0.27
False Positive Rate	0.14
False Negative Rate	0.48

4.1.3 LASSO regression

Next, I ran a LASSO regression. For the LASSO regression, Figure 10 shows the CV plot, Figure 11 shows the trace plot, and Table 6 shows the top 10 features with the largest standardized coefficients.

According to the CV plot, the LASSO regression selected 77 features if lambda is selected according to the one-standard-error rule. Like the ridge regression model, the LASSO regression found that older individuals, those without prior exposure to COVID, and those with higher educational attainments are less likely to be vaccine hesitant. Furthermore, the LASSO found that individuals who did not eat indoors in restaurants are also less likely to be vaccine hesitant. Interestingly, the LASSO also found that individuals with more children in their household are more likely to be vaccine hesitant.

The confusion matrix can be seen in Table 7 and the performance metrics for the LASSO regression can be seen in Table 8. The LASSO model has a misclassification rate of 0.26, which is higher than the logistic regression but slightly lower than the boosted. The LASSO has a higher false positive rate when compared to the ridge regression at 0.16 and it has the lowest false negative rate of all the regression models at 0.44.

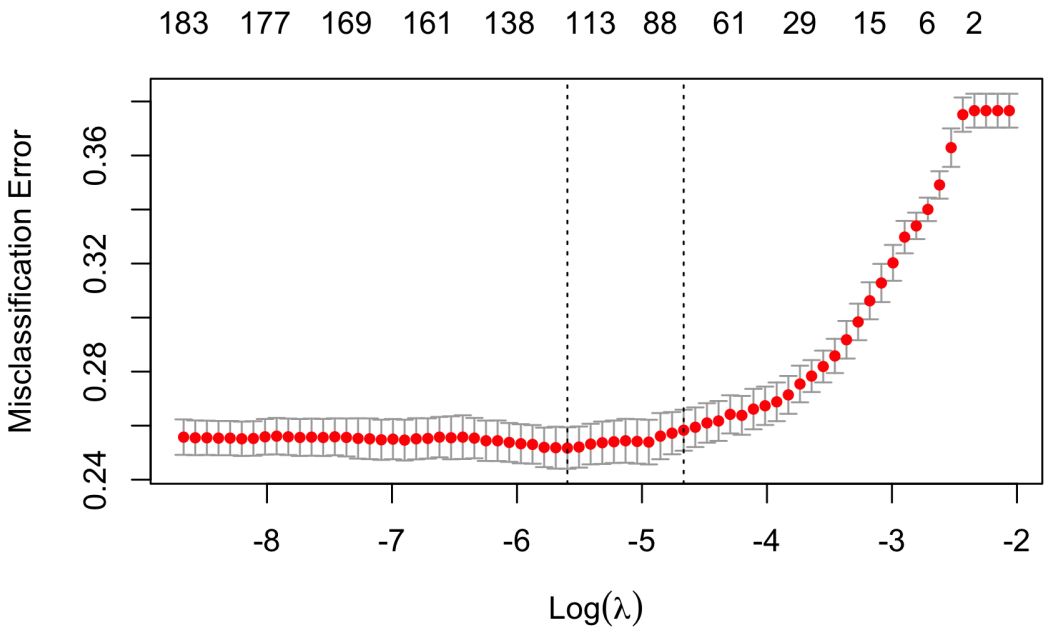

Figure 10: LASSO CV plot.

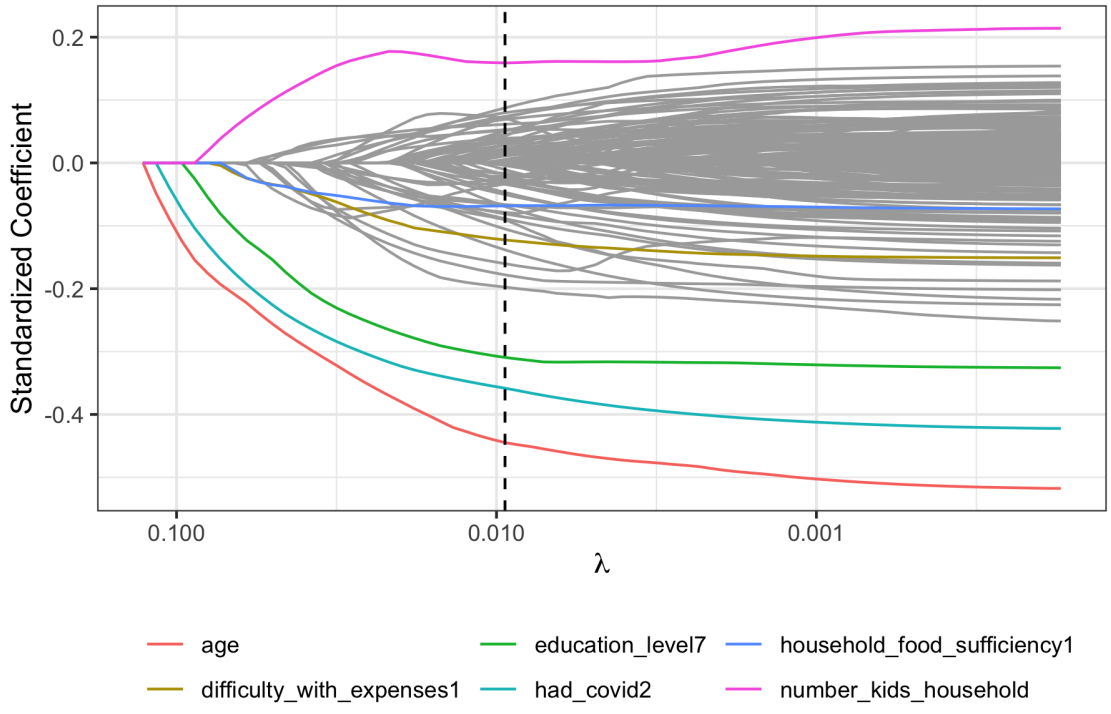

Figure 11: LASSO trace plot.

Table 6: Standardized coefficients for features in the lasso model based on the one-standard-error rule.

Feature	Coefficient
age	-0.44
had_covid2	-0.36
education_level7	-0.31
eating_indoors_restaurants2	-0.20
education_level6	-0.18
prescription_mental_health1	-0.16
number_kids_household	0.16
race3	-0.14
difficulty_with_expenses1	-0.12
income8	-0.11

Table 7: The confusion matrix for the LASSO regression

	0	1
0	1072	197
1	339	438

Table 8: The performance metrics of the LASSO regression

Metric	Classifier Performance
Misclassification Error	0.26
False Positive Rate	0.16
False Negative Rate	0.44

4.1.4 Elastic net regression

Finally, I ran a elastic net regression. Figure 12 shows the CV plot, Figure 13 shows the trace plot, and Table 9 shows the top 10 features with the largest standardized coefficients.

According to the CV plot, the elastic net regression selected roughly 91 features if lambda is selected according to the one-standard-error rule. Like the other regression models, the elastic net regression found that older individuals, those without prior exposure to COVID, and those with higher educational attainments are less likely to be vaccine hesitant. Like the LASSO regression, the elastic net regression also found that individuals who did not eat indoors in restaurants are also less likely to be vaccine hesitant, and it also found that individuals with more children in their household are more likely to be vaccine hesitant.

The confusion matrix can be seen in Table 10 and the performance metrics for the elastic net regression can be seen in Table 11. The elastic net model has a misclassification rate of 0.30, which is the highest of all the regression models. The elastic net regression has the lowest false positive rate out of all the models at 0.07 and it has the highest false negative rate of all the models at 0.67. The elastic net regression thus has the highest specificity.

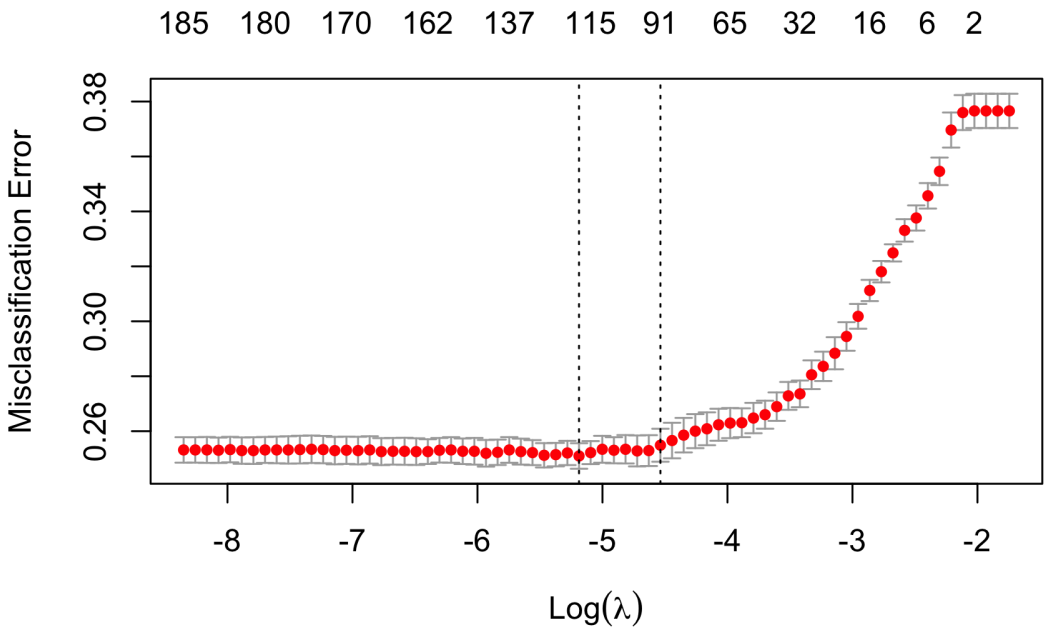

Figure 12: Elastic net CV plot.

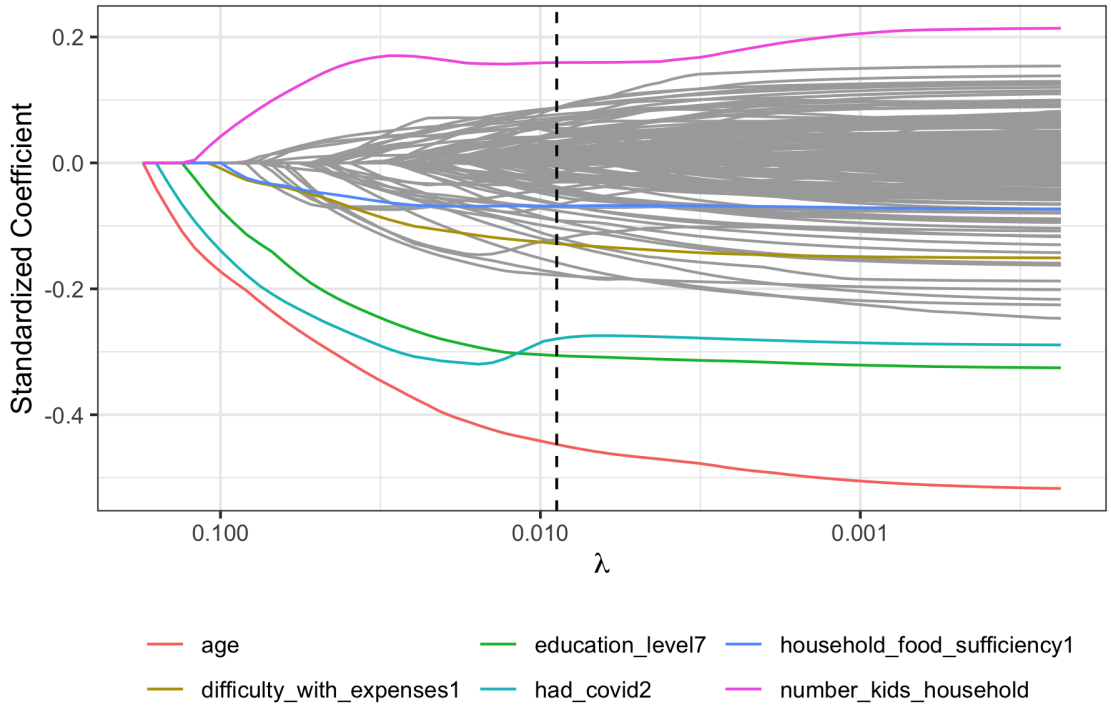

Figure 13: Elastic net trace plot.

Table 9: Standardized coefficients for features in the elastic net model based on the one-standard-error rule.

Feature	Coefficient
age	-0.45
education_level7	-0.31
had_covid2	-0.28
education_level6	-0.18
eating_indoors_restaurants2	-0.17
number_kids_household	0.16
race3	-0.16
difficulty_with_expenses1	-0.13
income8	-0.13
prescription_mental_health1	-0.12

Table 10: The confusion matrix for the elastic net regression

	0	1
0	1180	89
1	518	259

Table 11: The performance metrics of the elastic net regression

Metric	Classifier Performance
Misclassification Error	0.30
False Positive Rate	0.07
False Negative Rate	0.67

4.2 Tree-based methods

4.2.1 Classification trees

After fitting the regression models, I decided to fit two tree-based models: classification tree and boosted model.

For the classification tree, I fitted the deepest possible tree and pruned it to achieve the optimal tree. This optimal tree can be seen in Figure 14. There were 4 splits conducted that resulted in 5 terminal nodes. The sequence of splits that leads to the terminal node with the largest fraction of hesitant individuals is the following: 1. feature: education_level, split point: education_level = 6 or 7 (bachelor's or graduate degree), direction: frequency is not equal to 6 or 7 (i.e., right); 2. feature: age, split point: age >= 59 years, direction: age is < 59 years (i.e., right). Thus, the classification tree seems to suggest that individuals below 59 years of age with lower educational attainment are more likely to be vaccine hesitant. The splits that lead to the terminal node with the lowest fraction of hesitant individuals suggests that individuals with high educational attainment (bachelor's or graduate degree) and who did not have prior exposure to COVID are less likely to be vaccine hesitant.

The confusion matrix can be seen in Table 12 and the performance metrics for the classification tree can be seen in Table 13. The classification tree has a misclassification rate of 0.32, which is the highest of all the models. The classification tree has the highest false positive rate out of all models at 0.26 and a false negative rate of 0.40.

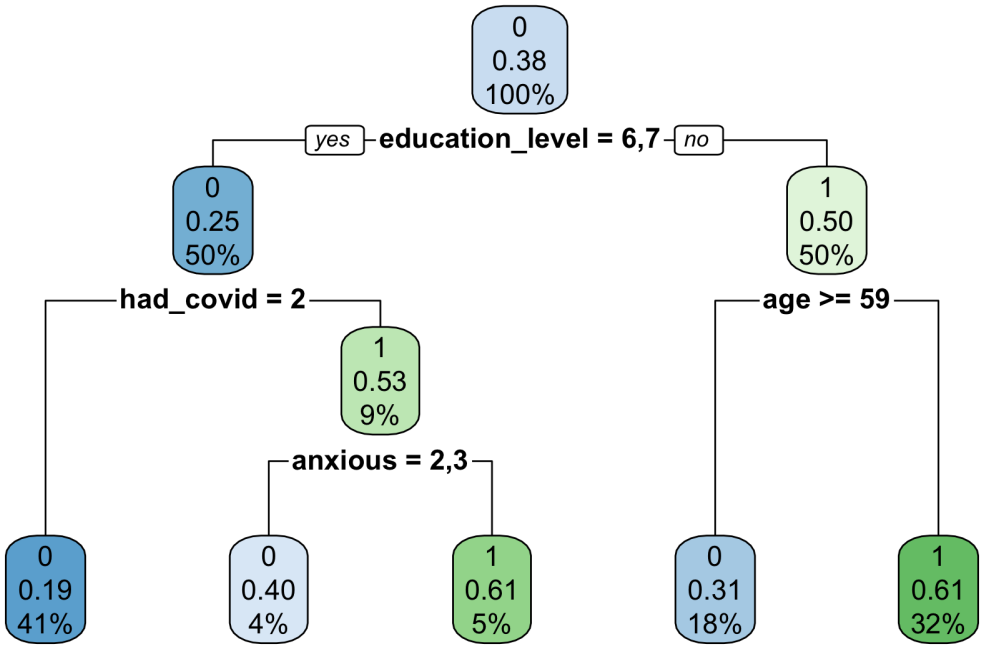

Figure 14: Classification tree plot.

Table 12: The confusion matrix for the classification tree

	0	1
0	935	334
1	314	463

Table 13: The performance metrics of the classification tree

Metric	Classifier Performance
Misclassification Error	0.32
False Positive Rate	0.26
False Negative Rate	0.40

4.2.2 Boosted

For my last prediction method, I decided to fit a boosted model, which aggregates the outputs of multiple classification trees and utilizes them to achieve good prediction performance.

To tune the boosted model, I experimented with the number of trees and a variety of interaction depths. I found that the optimal number of trees was between 200 to 400, and I tested the interaction depths of 1, 2, 3, and 4 (all else equal). As per the CV plot seen in Figure 15, I observed that the model with interaction depth 3 attains the minimum CV error with 954 trees.

With the tuned boosted model, I assessed variable importance through purity-based importance and partial

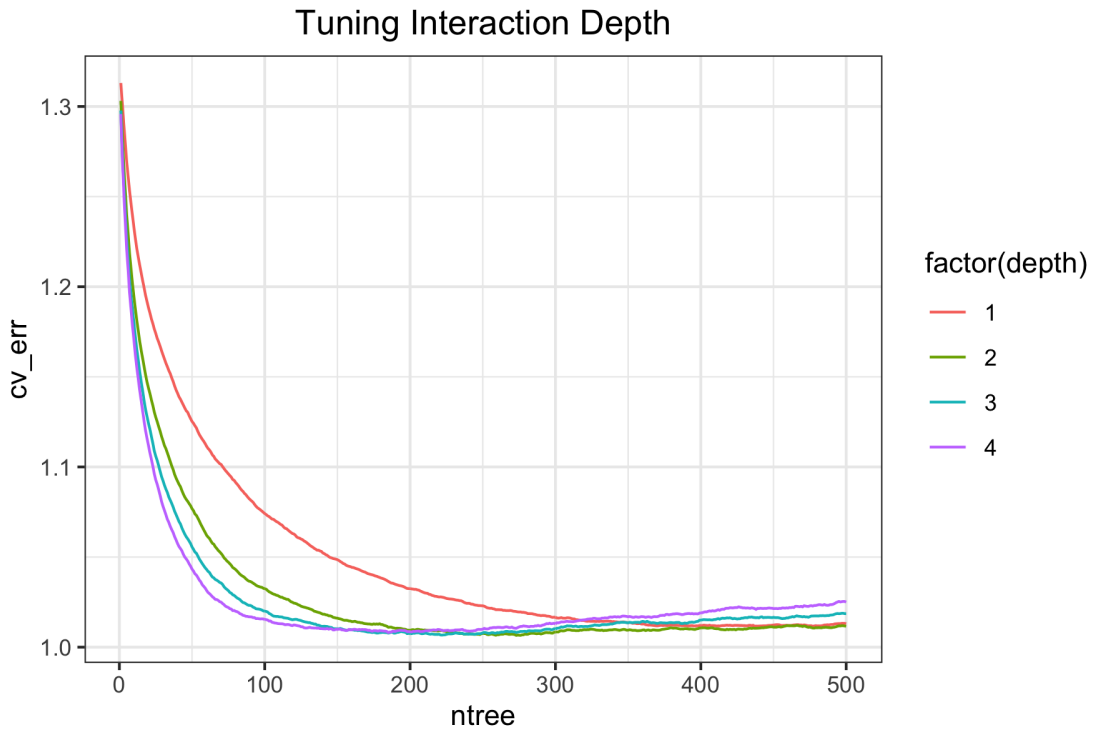

Figure 15: The CV error for the optimal boosted model

dependence plots. For purity-based importance, Table 14 shows the top 10 variables with the highest variable influence scores. The top variable is age, followed by education level, previous exposure to COVID, income, difficulty with expenses, eating indoors in restaurants, having anxiety, working from home, type of housing, and race.

Table 14: The variable influence scores for the boosted model

Variable	Relative Influence
age	17.06
education_level	13.91
had_covid	11.60
income	5.42
eating_indoors_restaurants	3.83
difficulty_with_expenses	3.75
anxious	3.20
building_type	2.74
race	2.58
prescription_mental_health	2.18

I then created partial dependence plots for the three most important variables: age, education level, and previous exposure to COVID. Figure 16 shows the partial dependence plot for age. From the plot, it can be seen that as age decreases, the likelihood of vaccine hesitancy increases. Figure 17 contains the partial dependence plot for education, and it shows that lower educational attainment is associated with vaccine hesitancy. Finally, Figure 18 shows the partial dependence plot for prior exposure to COVID. This plot shows that individuals who have tested positive for COVID or individuals who are unsure about testing positive for

COVID have a higher likelihood of vaccine hesitancy.

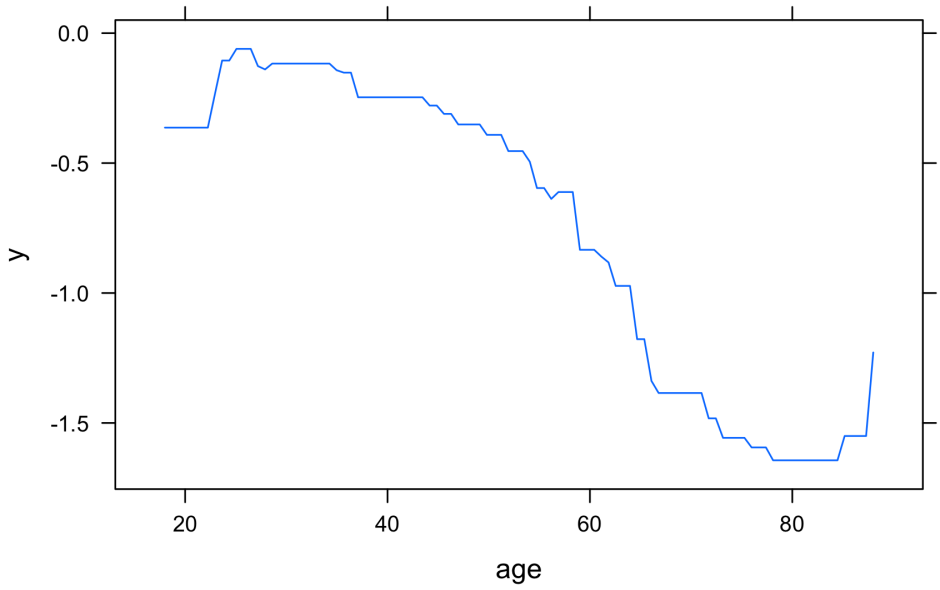

Figure 16: The partial dependence plot for age

The confusion matrix can be seen in Table 15 and the performance metrics for the boosted model can be seen in Table 16. The boosted model has a misclassification rate of 0.26, which is the highest of all the models. The boosted model has the highest false positive rate out of all models at 0.18 and a false negative rate of 0.39.

Table 15: The confusion matrix for the boosted model

	0	1
0	1044	225
1	305	472

Table 16: The performance metrics of the boosted model

Metric	Classifier Performance
Misclassification Error	0.26
False Positive Rate	0.18
False Negative Rate	0.39

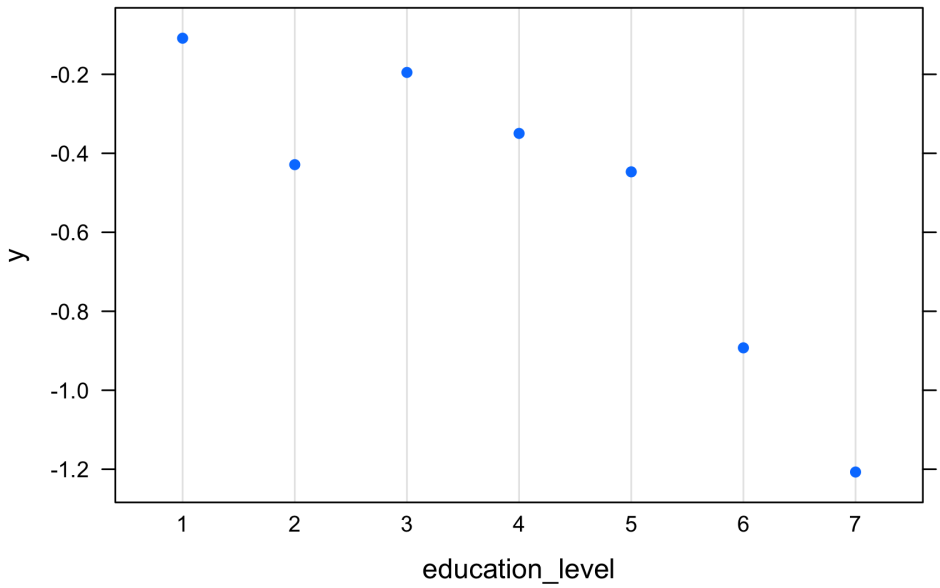

Figure 17: The partial dependence plot for education level

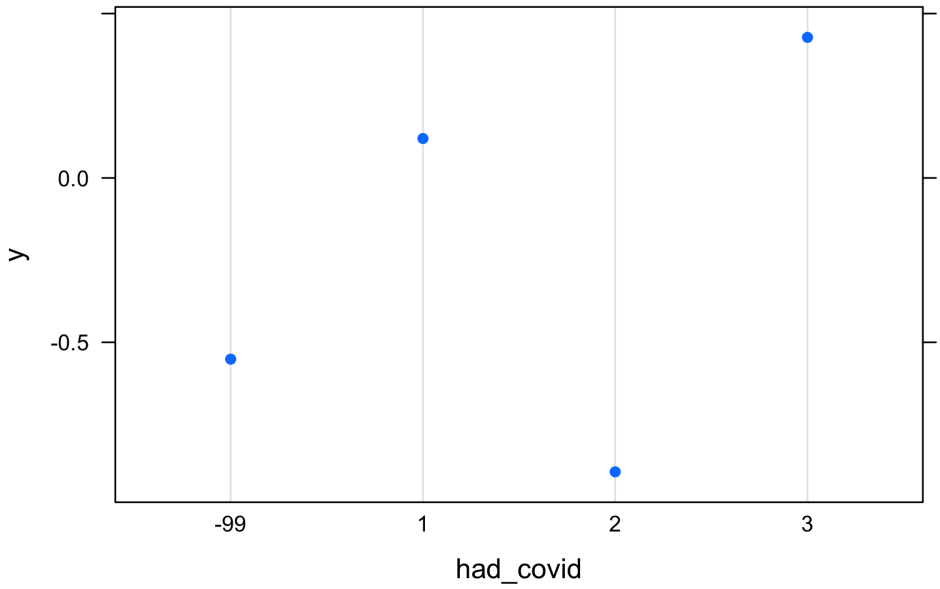

Figure 18: The partial dependence plot for prior COVID exposure

5 Conclusions

5.1 Method comparison

Table 17: Summary of performance for all models

Method	Misclassification Error	False Positive Rate	False Negative Rate
Logistic	0.25	0.17	0.39
Ridge	0.27	0.14	0.48
LASSO	0.26	0.16	0.44
Elastic Net	0.30	0.07	0.67
Classification Tree	0.32	0.26	0.40
Boosted	0.26	0.18	0.39

Table 17 shows the misclassification error, false positive rate, and false negative rate for all the methods considered. The logistic and the boosted model have the lowest misclassification errors at 0.25. I expected the boosted model to have a low misclassification error due to the boosted model's tendency to have high predictive accuracy due to its methodology of aggregating multiple trees. The superior performance of the logistic model when compared to the ridge, LASSO, and elastic net regressions suggest that bias can be a more dominant driver of test error than variance in this data. The classification tree performed the worst out of all the models, which was expected because although classification trees are easy to interpret, they can be unstable and often do not give the best prediction performance.

Even with these differences in misclassification error, the methods have significant overlap in their identification of important variables. Age was the most significant variable across nearly all the models, but other variables that were selected by more than one model included having a previous diagnosis of COVID, education level, race, eating indoors in restaurants, and the number of children in the individual's household.

5.2 Takeaways

My results point to a few key factors that public health officials should consider when aiming to improve the COVID-19 vaccination rate in the United States. The boosted model, which was tied for the strongest predictive performance, suggests that age is the most important variable in predicting an individual's vaccine hesitancy. This is corroborated by other well-performing models like the logistic regression, which also found that age is a significant predictor. This makes sense because much of the early research on COVID-19 showed that the disease is much more severe for older people than younger people. Because of this, younger people may then be less willing to get vaccinated. Education level is also highly important, as several models show that low educational levels are associated with a higher chance of vaccine hesitancy. Prior exposure to COVID was also an important variable in the boosted model, suggesting that those who already had COVID or were unsure if they had COVID were more likely to be vaccine hesitant. This may be because individuals who contracted COVID believe that they gained immunity from future infections, and thus are less inclined to get vaccinated. These variables are identified across all models, suggesting that these relationships are robust.

There are many possible ways that these results can be acted upon. Although younger people are less likely to be severely affected by COVID, they should still be vaccinated to build herd immunity. Thus, public health officials can adjust their messaging to better encourage younger people to get vaccinated. Another example is that since prior exposure for COVID was significant across multiple models, public health officials can work to better explain how one should still get vaccinated even after one contracts COVID to combat future variants of the disease.

That being said, with the constant evolution of the COVID-19 pandemic as well as the inherent complexity of vaccine hesitancy, I am hesitant to make any assertive claims about the true predictive capacity of any of

the top factors I identified. Nonetheless, these results can help inform policies directed toward improving vaccination rates across the US.

5.3 Limitations

5.3.1 Dataset limitations

Arguably the most significant limitation to my dataset is that vaccine hesitancy is an enormously complex and multifaceted issue. The dataset taken from the Household Pulse Survey does well at collecting base-level socioeconomic and demographic data, but there are a wide variety of factors that contribute to vaccine hesitancy that are not captured by this dataset. For example, the existence of pre-existing health issues is a significant cause for many to get vaccinated and it is a factor that is not captured by the Household Pulse Survey.

Another one of the major limitations of my dataset is that the data was collected through only seven weeks in 2021. Changes to vaccine-related policies have happened frequently throughout the pandemic, mostly due to the rapid development of such vaccines. These changes can include the approval of new vaccines, the introduction of government or employer vaccine mandates or news about the efficacy of the vaccine against new variants like Omnicon. Thus, as the pandemic fluctuates, the interpretation of the analysis may not be as applicable to specific points in time other than the period where the data was collected.

5.3.2 Analysis limitations

One of the significant limitations to my analysis was the major class imbalance in the raw dataset. The ratio of people who were considered not hesitant to people who were considered hesitant was around 10 to 1. I was able to reduce this class imbalance through downsampling, but this likely affected the predictive accuracy of my models due to the reduction in the number of samples being processed through the training dataset.

5.4 Follow-ups

To compensate for the limitations listed above, I believe that more extensive analysis should be done as more data on vaccine hesitancy is collected. There are several ongoing data collection efforts that seek to collect a wider range of factors that can impact vaccine hesitancy. One particular study of interest is the National Immunization Survey - Adult COVID Module (NIS) that is being conducted by the National Center for Immunization and Respiratory Diseases and the National Opinion Research Center at the University of Chicago[7]. The NIS survey aims to better understand the underlying reasons why people choose to get vaccinated or not get vaccinated. The NIS is collecting data on potential variables that are important but not present in the Household Pulse Survey, such as pre-existing health issues, ease of accessing vaccination services, and the vaccination statuses of the individual's friends and family. The NIS survey is set to be completed at the end of 2021, and I believe that it would be particularly interesting to conduct an in-depth analysis on the NIS data. Another analysis that would be interesting is looking at what factors drive vaccine hesitancy in other countries.

A Appendix: Descriptions of features

Below are the 62 features that I used for analysis. Words written in parentheses represent variable names. All variables are categorical unless otherwise noted.

Social factors:

- *Demographics*
  - Region (`region`): Indicates the individual's Census Region. 1 indicates Northeast (including the states of Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont), 2 indicates South (Alabama, Arkansas, Delaware, District of Columbia,

[7] https://www.cdc.gov/vaccines/imz-managers/nis/about.html

Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia), 3 indicates Midwest (Indiana, Illinois, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin), and 4 indicates West (Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, and Wyoming)
  - Race (`race`): Denotes race. 1 indicates White, 2 indicates Black, 3 indicates Asian, 4 indicates any other race or race in combination.
  - Hispanic (`hispanic`): Indicator of Hispanic origin. 1 indicates not of Hispanic origin, 2 indicates of Hispanic origin.
  - Education_level (`education_level`): Denotes educational attainment. 1 indicates less than high school, 2 indicates some high school, 3 indicates high school graduate or equivalent (for example GED), 4 indicates some college but degree not received or is in progress, 5 indicates associate's degree (for example AA, AS), 6 indicates bachelor's degree (for example BA, BS, AB), 7 indicates graduate degree (for example master's, professional, doctorate).
  - Marital status (`marital_status`): Denotes marital status. 1 indicates married, 2 indicates widowed, 3 indicates divorced, 4 indicates separated, 5 indicates never married, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Birth gender (`birth_gender`): Denotes gender at birth. 1 indicates male, 2 indicates female.
  - Described gender (`described_gender`): Denotes current gender identity. 1 indicates male, 2 indicates female, 3 indicates transgender, 4 indicates none of the above, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Sexual orientation (`sexual_orientation`): Sexual orientation. 1 indicates gay or lesbian, 2 indicates heterosexual, 3 is bisexual, 4 is something else, 5 is doesn't know, -99 is question seen but category not selected, -88 is missing or did not report.
  - Age (`age`): Denotes age of individual. This is a numerical feature.

Economic factors

- *Household*
  - Kids (`kids`): Indicates having kids in household (defined as 18 years or younger).
  - Number of kids (`number_kids_household`): Number of kids in household (defined as 18 years or younger). This is a numerical feature.
  - Number of adults (`number_adults_household`): Number of adults in household (defined as 18 years or older). This is a numerical feature.
  - Negative childcare impact (`pandemic_child_impact`): Indicator of negative effect on childcare due to pandemic. Childcare impact includes taking unpaid leave to care for children, using vacation, sick days, or paid leave in order to care for children, cutting work hours in order to care for children, leaving a job in order to care for the children, not looking for a job in order to care for the children, supervising one or more children while working, and so on. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
- *Employment*
  - Recent work loss (`recent_work_loss`): Indicator of recent household job loss within the past 4 weeks. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Employment (`employment`): Indicator of any work done for either pay or profit over the past 7 days. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Armed forces service (`active_duty`): Indicator of the indvidual or their spouse serving in the U.S. Armed Forces (active duty, Reserve, National Guard). 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Essential worker (`essential_worker`): Indicates if individual is essential worker. Essential work are defined as working in hospitals, nursing and residential healthcare facilities, ambulatory healthcare, social service, education, first response, death care, correctional facility, food and beverage store, agriculture, forestry, fishing, hunting, manufacturing facilities, public transit, postal service, and so on. 1 indicates yes, 2 indicates no, -99 indicates question seen but category not

selected, -88 indicates missing or did not report.

- *Financial Stability*
  - Child Tax Credit (`child_tax_credit`): Indicates the receipt of Child Tax Credit (CTC) payment(s) in the last 4 weeks. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Difficulty with expenses (`difficulty_with_expenses`): Indicates difficulty with paying for usual household expenses, including but not limited to food, rent or mortgage, car payments, medical expenses, student loans, and so on. 1 indicates not at all difficult, 2 indicates a little difficult, 3 indicates somewhat difficult, 4 indicates very difficult, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Household food sufficiency (`household_food_sufficiency`): Indicates household food sufficiency over the past 7 days. 1 indicates enough of the kinds of food (I/we) wanted to eat, 2 indicates enough, but not always the kinds of food (I/we) wanted to eat, 3 indicates sometimes not enough to eat, 4 indicates often not enough to eat, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Received free food (`household_free_food`): Indicates if the individual or household recieved free groceries from a food pantry, food bank, church, or other place that helps with free food. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Received SNAP benefits (`snap_receipt`): Indicates if the individual or household received benefits from the Supplemental Nutrition Assistance Program (SNAP) or the Food Stamp Program. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Difficult paying energy bill (`energy_pay_trouble`): Indicates if the household has trouble paying the energy bill. 1 indicates almost every month, 2 indicates some months, 3 indicates 1 or 2 months, 4 indicates never, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Unsafe housing temperatures (`housing_unsafe_temp`): Indicates if the household keeps home at a temperature that is unsafe or unhealthy. 1 indicates almost every month, 2 indicates some months, 3 indicates 1 or 2 months, 4 indicates never, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Income (`income`): Denotes annual household income level. 1 indicates less than `$25,000`, 2 indicates between `$25,000 - $34,999`, 3 indicates `$35,000 - $49,999`, 4 indicates `$50,000 - $74,999`, 5 indicates `$75,000 - $99,999`, 6 indicates `$100,000 - $149,999`, 7 indicates `$150,000 - $199,999`, 8 indicates `$200,000` and above, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
- *Education*
  - Summer educational catch-up activities (`summer_educational_catch_up`): Indicator of children having to attend summer educational catch-up activities. 1 indicates yes, 2 indicates no, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Changes to post-secondary education plan (`post_sec_edu_plan_changes`): Indicator of changes made to post-secondary education plan including canceled classes, classes in different formats, fewer classes being taken, more classes being taken, classes being taken from a different institution, changes to degree, and so on. 1 indicates yes, 2 indicates no, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Children enrolled in public school (`children_public_school`): Indicator that at least one child is enrolled in public school. 1 indicates yes, 2 indicates no, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Children enrolled in private school (`children_private_school`): Indicator that at least one child is enrolled in private school. 1 indicates yes, 2 indicates no, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Children homeschooled (`children_homeschool`): Indicator that at least one child is homeschooled. 1 indicates yes, 2 indicates no, -99 indicates question seen but category not selected, -88 indicates missing or did not report.

  - Post secondary classes (`post_secondary_classes`): Indicator that at least one member of household was planning to take post secondary classes. 1 indicates yes, 2 indicates no, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
- *Housing*
  - Housing owned or rented (`housing_owned_rented`): Indicates if the individuals' housing is owned or rented. 1 indicates owned by individual or someone in the household free and clear, 2 indicates owned by individual or someone in the household with a mortgage or loan (including home equity loans), 3 indicates rented, 4 indicates occupied without payment of rent, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Housing building type (`building_type`): Indicates the individuals' housing type. 1 indicates a mobile home, 2 indicates a one-family house detached from any other house, 3 indicates Aa one-family house attached to one or more houses, 4 indicates a building with 2 apartments, 5 indicates a building with 3 or 4 apartments, 6 indicates a building with 5 or more apartments, 7 indicates a boat, RV, van, etc. -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Caught up on rent (`rent_caught_up`): Indicates if the household is currently caught up on rent payments. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Caught up on mortgage (`mortgage_caught_up`): Indicates if the household is currently caught up on mortgage payments. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Housing pay confidence (`housing_pay_confidence`): Denotes confidence in ability to make next mortgage or rent payment on time. 1 indicates not at all confident, 2 indicates slightly confident, 3 indicates moderately confident, 4 indicates highly confident, 5 indicates payment is/will be deferred, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Application for emergency rental assistance (`rent_assistance_apply`): Indicates if the household applied for emergency. 1 indicates the household applied and received assistance, 2 indicates the household applied and is waiting for a response, 3 indicates the household applied and the application was denied, 4 indicates the household did not apply, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Eviction likelihood (`eviction_likelihood`): Indicates likelihood of eviction within the next two months. 1 indicates very likely, 2 indicates somewhat likely, 3 indicates not very likely, 4 indicates not likely at all, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Foreclosure likelihood (`foreclose_likelihood`): Indicates likelihood of eviction within the next two months. 1 indicates very likely, 2 indicates somewhat likely, 3 indicates not very likely, 4 indicates not likely at all, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
- *Health insurance*
  - Private health insurance (`private_health_insurance`): Indicates if the individual has private health insurance. 1 indicates yes, 2 indicates no, 3 indicates missing or did not report.
  - Public health insurance (`public_health_insurance`): Indicates if the individual has public health insurance. 1 indicates yes, 2 indicates no, 3 indicates missing or did not report.

Health factors:

- *Vaccination Status*
  - Vaccine hesitancy (`vaccine_hesitancy`): Indicator of intention on getting vaccine. 1 indicates will definitely get a vaccine, 2 indicates will probably get a vaccine, 3 indicates unsure about getting a vaccine, 4 indicates will probably not get a vaccine, 5 indicates will definitely not get a vaccine, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Had COVID-19 (`had_covid`): Indicator of doctor or provider diagnosis of COVID. 1 indicates yes, 2 indicates no, 3 indicates not sure, -99 indicates question seen but category not selected, -88

indicates missing or did not report.

- *Emotional well-being*
  - Anxiety (`anxious`): Denotes the individual's frequency of anxiety over previous 2 weeks. 1 indicates no anxiety, 2 indicates having anxiety for several days, 3 indicates having anxiety for more than half the days, 4 indicates having anxiety nearly every day, -99 indicates question seen but category not selected, -88 indicates missing or id not report.
  - Worry (`worry`): Denotes the individual's frequency of worry over previous 2 weeks. 1 indicates not being worried, 2 indicates being worried for several days, 3 indicates being worried for more than half the days, 4 indicates being worried nearly every day, -99 indicates question seen but category not selected, -88 indicates missing or id not report.
  - Losing interest (`interest`): Denotes the individual's frequency of having little interest in things over previous 2 weeks. 1 indicates no loss of interest, 2 indicates having little interest for several days, 3 indicates having little interest for more than half the days, 4 indicates having little interest nearly every day, -99 indicates question seen but category not selected, -88 indicates missing or id not report.
  - Depression (`depressed`): Denotes the individual's frequency of feeling depressed over previous 2 weeks. 1 indicates no feelings of depression, 2 indicates having feelings of depression for several days, 3 indicates having feelings of depression for more than half the days, 4 indicates having feelings of depression nearly every day, -99 indicates question seen but category not selected, -88 indicates missing or id not report.
- *Mental health*
  - Prescription medication for mental health (`prescription_mental_health`): Usage of prescription medication within the past 4 weeks. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Mental health services (`mental_health_services`): Indicator of receiving counseling or therapy from a mental health professional within the past 4 weeks. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - No access to mental health services (`mental_health_not_get`): Indicator of needing counseling or therapy from a mental health professional within the past 4 weeks but not receiving it. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
- *Physical health*
  - Missed checkups for child (`child_missed_checkups`): Children in household missed or delayed preventive check-ups within the last 12 months. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Seeing limitations (`seeing_limitations`): Indicator of having difficulty seeing, even when wearing glasses or contacts. 1 indicates no difficulty, 2 indicates some difficulty, 3 indicates a lot of difficulty, 4 indicates cannot do at all, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Hearing limitations (`hearing_limitations`): Indicator of having difficulty hearing, even when using a hearing aid. 1 indicates no difficulty, 2 indicates some difficulty, 3 indicates a lot of difficulty, 4 indicates cannot do at all, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Remembering limitations (`remembering_limitations`): Indicator of having difficulty remembering or concentrating. 1 indicates no difficulty, 2 indicates some difficulty, 3 indicates a lot of difficulty, 4 indicates cannot do at all, -99 indicates question seen but category not selected, -88 indicates missing or did not report.
  - Mobility limitations (`mobility_limitations`): Indicator of having difficulty walking or climbing stairs. 1 indicates no difficulty, 2 indicates some difficulty, 3 indicates a lot of difficulty, 4 indicates cannot do at all, -99 indicates question seen but category not selected, -88 indicates missing or did not report.

Lifestyle factors:

- *Work*
  - Working onsite (`work_onsite`): Indicator of working onsite within the past 7 days by the individual

or anyone in the household. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Working from home (`work_from_home`): Indicator of working from home within the past 7 days by the individual or anyone in the household. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
- *Personal*
  - In-store shopping (`in_store_shopping`): Indicator of shopping in-store within the past 7 days by the individual or anyone in the household. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Eating indoors at restaurants (`eating_indoors_restaurants`): Indicator of eating indoors at restaurants within the past 7 days by the individual or anyone in the household. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - In-home housekeeping or caregiving services (`in_home_housekeeping`): Indicator of having in-person housekeeping or caregiving services within the past 7 days by the individual or anyone in the household. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
- *Health*
  - In-person medical appointments (`in_person_medical_appointments`): Indicator of having in-person medical appointments within the past 7 days by the individual or anyone in the household. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Personal usage of telehealth (`telehealth_personal`): Personal use of telehealth or telemedicine within the past 4 weeks. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.
  - Child's usage of telehealth (`telehealth_child`): Child's use of telehealth or telemedicine within the past 4 weeks. 1 indicates yes, 2 indicates no, -99 is question seen but category not selected, -88 is missing or did not report.