Predicting COVID vaccine uptake using XGBoost and SHAP

James Farrell
Florida Southern College

Abstract

This paper explores the relationship between COVID vaccine uptake and individual demographic characteristics using an XGBoost model based on survey results from the Federal Reserve's 2021 Survey of Household Economic Decisionmaking (SHED). This paper walks through the tuning steps and model progress to achieve a predicted accuracy of 72%, with sensitivity and specificity of 75% and 59%, respectively. These results are comparable to other studies which used the XGBoost model to focus on county-level vaccine uptake. The results were further explored using SHAP to extract variable importance. Greater age and higher levels of income and education all contributed positively to the likelihood of vaccine uptake. One surprising result was that the presence of older parents living with the survey respondents was not an important variable in predicting vaccine uptake.

Keywords: COVID Vaccine Uptake; XGBoost; SHAP; SHED; Predicted Accuracy; Demographic Characteristics

Introduction

This paper investigates the factors which drove the individual decision on whether to take the COVID-19 vaccine. Following the COVID pandemic and rapid development of vaccines, there was a significant resistance to vaccine acceptance in the US, among other countries (Sallam, 2021). While, on the surface, much of the difference appeared to be across political lines in the U.S. (Shmerling, 2021), it is worthwhile to take a deeper look to explore a model which may more accurately predict vaccine acceptance. This is an important problem to understand going forward as future pandemics may require a different approach and understanding the likelihood of a population accepting the vaccine may improve vaccine rates.

To investigate this problem, the paper will utilize the 2021 Survey of Household Economic Decisionmaking (SHED) dataset from the Federal Reserve. This data provides the result of the survey where participants were asked about whether they received the COVID vaccine along with a rich set of demographic data. The Data Overview section of the paper will take a deeper dive, but, overall, this data provides an opportunity to analyze the contributing factors to this question.

To analyze the data, an XGBoost model was used and SHAP results are presented in the results and discussion section. This model was tuned to balance the response variable and tune other hyper-tuning elements, this is described in detail in the Methods section.

Literature Review

This question has been addressed by prior literature using various approaches and datasets. Cheong et. al (2021) took a similar approach and used an XGBoost model to study vaccine acceptance at the county level. Overall, they found several socioeconomic variables, including location, education, ethnicity, income, and household access to internet as the most important factors in predicting vaccine acceptance. Their model showed meaningful success with a 62% accuracy.

Schmitz et. al (2022) took a different approach and studied individual level decisions using a cross-sectional and longitudinal study of Belgian participants to understand the motivation (or lack thereof) to get vaccinated. Their study focused more on the role of risk of infection perceptions and feelings of autonomy rather than features/characteristics of the study participants. While this is an interesting result, studying individual opinions about the disease seems like it would be naturally very predictive of their choices, however, and offer little insight to predicting who would be resistant to vaccine acceptance without already knowing their opinions. Their study was accomplished with SEM and confirmatory factor analysis.

Another paper used a longitudinal analysis to assess hesitancy attitudes and uptake with a logistic model (Latkin et. al, 2022). Their paper found that opinions of friends/family, uncertainty about whom to believe, and uncertainty about shortcuts, along with certain demographic variables (political preference, gender, education, and income) were all impactful factors in vaccine acceptance.

This paper should present a unique approach to addressing this question. Similar to Cheong et. al (2021), this paper will use an XGBoost model, but rather than focus on county-level outcomes, this paper will focus on individual outcomes. This paper will also leave out personal opinions about vaccines as predictors as the goal is to identify the features/characteristics which may be known without a survey about COVID to predict whether

an individual would accept the vaccine. While Schmitz et. al (2022) and Latkin et. al, (2022) provided interesting insights into the role that preferences/psychology may play into vaccine acceptance, this paper will take a different approach by limiting the explanatory variables to non-opinion features. Overall, the focus on individual-level data and methodology should enable this analysis to address the COVID vaccine acceptance question in a new way.

Data Overview

For the 2021 survey year the Federal Reserve included questions on COVID vaccine uptake and opinions (safety, necessity etc.) along with their usual set of questions in their annual SHED data (Federal Reserve Board, 2021). The full dataset contains nearly 12,000 observations and over 1,100 variables. This analysis will focus on key explanatory variables that cover individual demographics (age, income, education, marital status by gender, and race) along with household makeup (presence of children, partners, adult parents) and location (state and region). While this dataset contains variables which address opinions about vaccine safety and efficacy, these variables will be excluded as the focus of the analysis is on the features/characteristics that may predict vaccine acceptance.

Some of the key variables are summarized in Tables 1 – 4 (Appendix), categorized by Vaccinate and Not Vaccinated. From the means of the data, we can see that Vaccine uptake increased with age, income and education, but there were only slight differences across race.

To better see the relationships, Age and Income are represented graphically in Figure 1 (Appendix) and Figure 2 (Appendix). From the Figure 1, we can see the pattern of younger respondents being less likely to be vaccinated. This is an expected relationship as older respondents were generally considered more at risk of serious complications. From Figure 2 we can see another strong relationship, with higher income individuals more likely to be vaccinated than lower income individuals. There is likely a more complex relationship here as there is meaningful correlation between age and income.

Methods

As mentioned in the Introduction section, XGBoost was used to create a predictive model for classifying respondents into Vaccinated and Not Vaccinated classes. The XGBoost model was chosen because it allows for complex relationships between independent variables, is well suited for classification problems, allows for weight-balancing to account for the natural imbalance in the data, and hyper-tuning to create well-, but not over-fitted, model that balances accuracy, sensitivity, and specificity (Chen and Guestrin, 2016). The complexity of this model has the drawbacks of limited interpretability of individual results, e.g., you cannot visualize a decision tree. However, combining this with SHAP analysis regains some of the interpretability.

The full work of the analysis and tuning process is included in the R-code, but to summarize the steps taken:

1. Initial model to get a baseline set of results:
    a. Accuracy: 0.8084
    b. Sensitivity: 0.9403
    c. Specificity: 0.2153

2. The first tuning step was rebalancing the data, resulted in:
    a. Accuracy: 0.7411
    b. Sensitivity: 0.8219
    c. Specificity: 0.3773

3. The second tuning step was hyper-tuning the max_depth_vals, min_child_weight, gamma, subsample, colsample_bytree and finally eta, this resulted in choices of:
    - eta = 0.3
    - max_depth = 7
    - min_child_weight = 5
    - gamma = 0.1
    - subsample = 0.7
    - colsample_bytree = 0.6

    And final results of:
    a. Accuracy: 0.7175
    b. Sensitivity: 0.7458
    c. Specificity: 0.5903

Results and Discussion

The tuning process led to meaningful improvement in the balance of specificity, sensitivity, and accuracy. The balance of the specificity and sensitivity were prioritized over the accuracy of the model as it is important to accurately predict both sides of the vaccine acceptance decision. The final confusion matrix and AUC are included as Figures 3 and 4 (Appendix). The overall accuracy of 72% is a strong result when compared to Cheong et. al (2021), which achieved an accuracy of 62% when looking for results at the county-level.

To understand which characteristics are important to predicting COVID vaccine acceptance Shapley Additive explanation, or SHAP, was used (Lundberg, 2017). The results of the SHAP reveal some interesting contributing factors in Figure 5 (Appendix). The most important factors that increased the probability of being vaccinated were the upper age categories (65 – 74 and 75+), higher levels of education (Bachelor's, Master's) and higher levels of income ($100k - $150K, $150K - $200K and $200k+). The most important factors for decreasing the probability of being vaccinated are the presence of children in the household (L0_b1), lower levels of education (HS/GED, less than HS) and lower age categories (25 – 34). These results are in line with findings from other analyses at both the individual and aggregate levels and offer a nice confirmation of existing expectations.

One set of features that this analysis covered that prior papers did not cover were on household makeup… whether you lived with a partner, young children, adult children, your parents (L0_a through L0_d). I had initially expected that living with your parents would increase your likelihood to get vaccinated. This was based on the idea that adults living with older, vulnerable parents would get vaccinated. The L0_d variable, however, did not rise to the level of importance I expected. Perhaps this was caused by the mix of younger adults still living with relatively younger parents, along with older adults living with even older parents.

Conclusion and Future Work:

Overall, this was an interesting exercise and XGBoost, once fully tuned, provided a reasonably predictive and well-balanced model for COVID vaccine acceptance. The methodology is comparable to Cheong et. al (2021), with a focus on individual choices rather than county-level aggregates, with the final results achieving solid accuracy and balance. The variables of importance lined-up with the prior expectations that higher income, more educated, and older individuals were more likely to get vaccinated. To improve overall vaccine acceptance, improving messaging targeted at younger and less educated individuals may help close the gap should a future outbreak occur.

Future research into this topic could include running separate analysis on younger and older households to see whether the household makeup variables, particularly L0_d (living with parents), become meaningful factors in the vaccination decision.

Statements and Declarations

Competing Interests Statement: The author has no competing interests either directly or indirectly related to this paper.

Data Availability Statement: The dataset used in this study is based on publicly available data from the Federal Reserve's SHED survey.

References

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Cheong Q, Au-yeung M, Quon S, Concepcion K, Kong J, Predictive Modeling of Vaccination Uptake in US Counties: A Machine Learning–Based Approach, *J Med Internet Res 2021;23(11):e33231*, URL: https://www.jmir.org/2021/11/e33231, DOI: 10.2196/33231

Latkin, C., Dayton, L., Miller, J., Yi, G., Balaban, A., Boodram, B., Uzzi, M., & Falade-Nwulia, O. (2022). A longitudinal study of vaccine hesitancy attitudes and social influence as predictors of COVID-19 vaccine uptake in the US. *Human Vaccines & Immunotherapeutics, 18*(5). https://doi.org/10.1080/21645515.2022.2043102

Federal Reserve Board. 2021. Survey of Household Economics and Decisionmaking (SHED) [Data set]. Federal Reserve Board. https://www.federalreserve.gov/consumerscommunities/shed_data.htm

Lundberg, S., Lee, S. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*. https://doi.org/10.48550/arXiv.1705.07874

Mathias Schmitz, Olivier Luminet, Olivier Klein, Sofie Morbée, Omer Van den Bergh, Pascaline Van Oost, oachim Waterschoot, Vincent Yzerbyt, Maarten Vansteenkiste, Predicting vaccine uptake during COVID-19 crisis: A motivational approach, Vaccine, Volume 40, Issue 2, 2022, Pages 288-297, ISSN 0264-410X, https://doi.org/10.1016/j.vaccine.2021.11.068. (https://www.sciencedirect.com/science/article/pii/S0264410X21015425)

Sallam, M. COVID-19 Vaccine Hesitancy Worldwide: A Concise Systematic Review of Vaccine Acceptance Rates. *Vaccines* 2021, *9*, 160. https://doi.org/10.3390/vaccines9020160

Shmerling, Robert H. (2021, August 25). Unvaccinated and misunderstood: Let's talk. Harvard Health Blog. https://www.health.harvard.edu/blog/unvaccinated-and-misunderstood-lets-talk-202108252580

Appendix

Table 1

Characteristic	Not Vaccinated[1]	Vaccinated[1]
Age		
18 - 24	136 (6.3%)	429 (4.4%)
25 - 34	530 (24.4%)	1,416 (14.6%)
35 - 44	464 (21.4%)	1,318 (13.6%)
45 - 54	356 (16.4%)	1,355 (14.0%)
55 - 64	393 (18.1%)	2,098 (21.6%)
65 - 74	211 (9.7%)	2,061 (21.2%)
75+	82 (3.8%)	1,025 (10.6%)

[1] n (%)

Table 2

Characteristic	Not Vaccinated[1]	Vaccinated[1]
Race		
White	1,520 (70.0%)	6,872 (70.8%)
Black	238 (11.0%)	949 (9.8%)
Hispanic	288 (13.3%)	1,116 (11.5%)
Asian	29 (1.3%)	425 (4.4%)
Other	97 (4.5%)	340 (3.5%)

[1] n (%)

Table 2

Characteristic	Not Vaccinated[1]	Vaccinated[1]
Education		
Less than HS	198 (9.1%)	304 (3.1%)
HS or GED	661 (30.4%)	1,651 (17.0%)
Some College	475 (21.9%)	1,675 (17.3%)
Cert or Technical	125 (5.8%)	480 (4.9%)
Associates	227 (10.5%)	846 (8.7%)
Bachelor's	343 (15.8%)	2,619 (27.0%)
Master's	103 (4.7%)	1,347 (13.9%)
Professional	28 (1.3%)	483 (5.0%)
Doctoral	12 (0.6%)	297 (3.1%)

[1] n (%)

Table 1

Characteristic	Not Vaccinated[1]	Vaccinated[1]
Income		
$0	214 (9.9%)	357 (3.7%)
$1 to $4,999	190 (8.7%)	407 (4.2%)
$5,000 to $14,999	199 (9.2%)	490 (5.1%)
15,000 to $24,999	190 (8.7%)	626 (6.5%)
$25,000 to $39,999	263 (12.1%)	860 (8.9%)
$40,000 to $49,999	186 (8.6%)	673 (6.9%)
$50,000 to $74,999	312 (14.4%)	1,478 (15.2%)
$75,000 to $99,999	233 (10.7%)	1,163 (12.0%)
$100,000 to $149,999	223 (10.3%)	1,636 (16.9%)
$150,000 to $199,999	104 (4.8%)	1,015 (10.5%)
$200,000 or higher	58 (2.7%)	997 (10.3%)

[1] n (%)

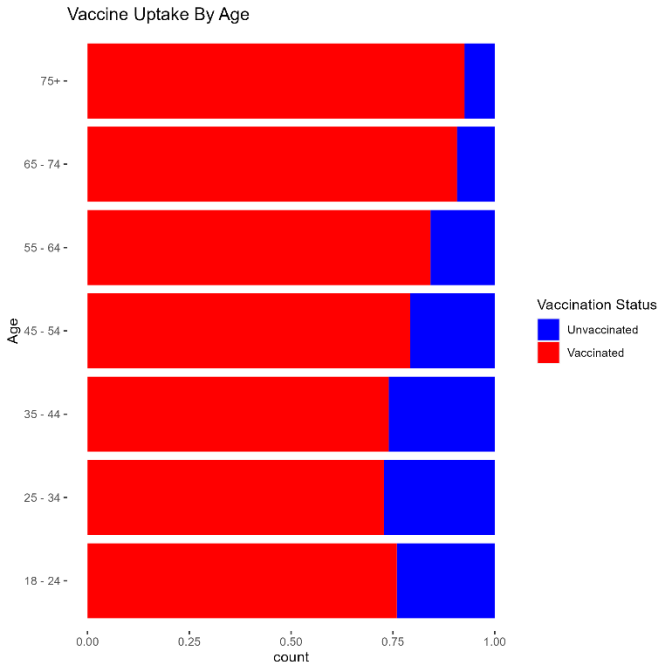

Figure 1

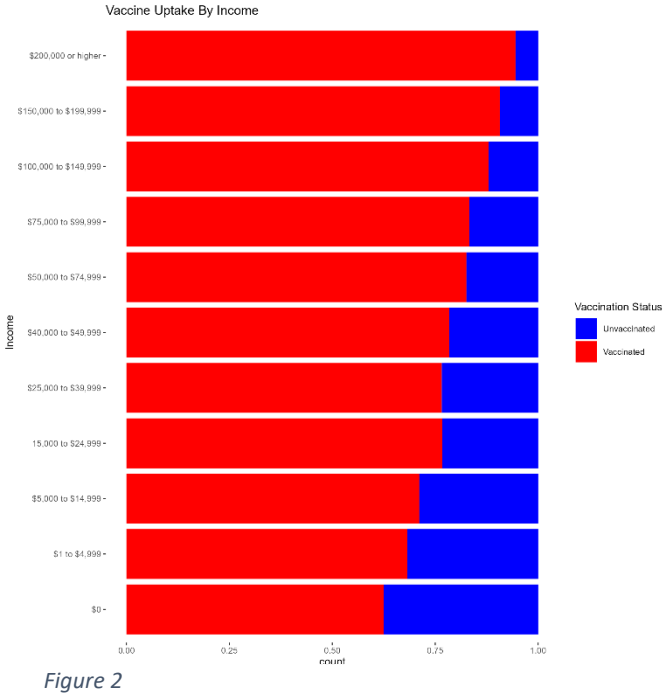

Figure 2

```
Confusion Matrix and Statistics


boost_pred_class    0    1
               0  255  494
               1  177 1449

               Accuracy : 0.7175
                 95% CI : (0.6989, 0.7355)
    No Information Rate : 0.8181
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2614

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7458
            Specificity : 0.5903
         Pos Pred Value : 0.8911
         Neg Pred Value : 0.3405
             Prevalence : 0.8181
         Detection Rate : 0.6101
   Detection Prevalence : 0.6846
      Balanced Accuracy : 0.6680

       'Positive' Class : 1
```

Figure 3

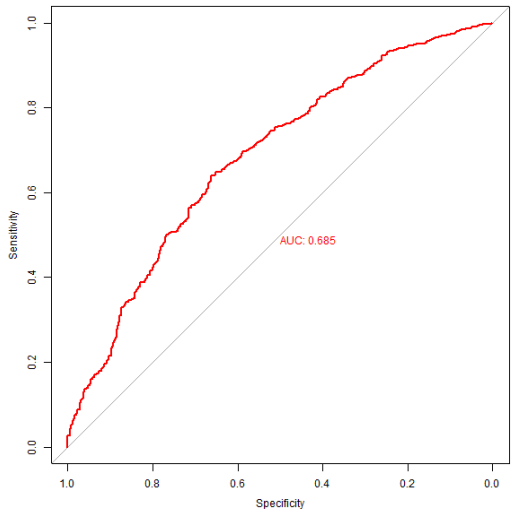

Figure 4

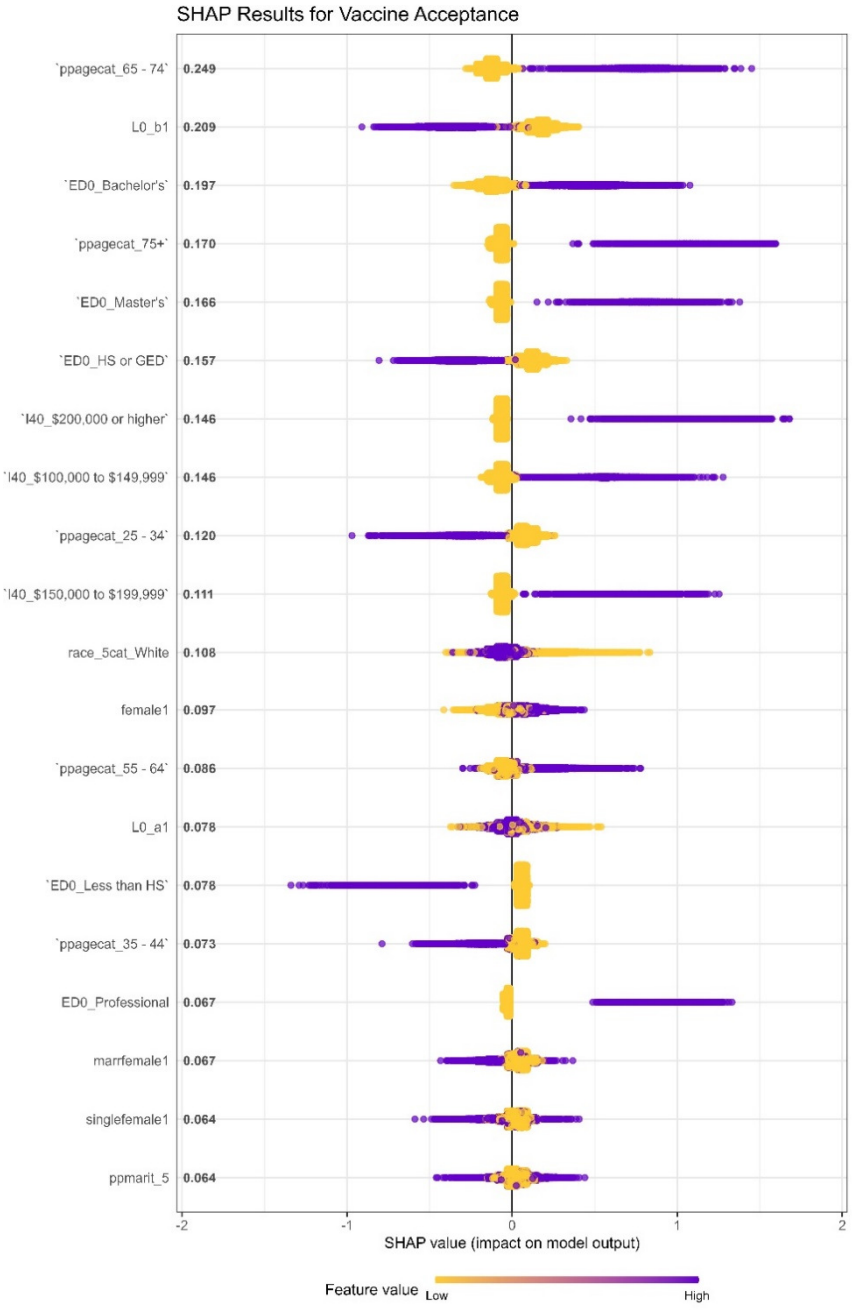
SHAP Results for Vaccine Acceptance
`ppagecat_65 - 74` 0.249
L0_b1 0.209
`ED0_Bachelor's` 0.197
`ppagecat_75+` 0.170
`ED0_Master's` 0.166
`ED0_HS or GED` 0.157
`I40_$200,000 or higher` 0.146
`I40_$100,000 to $149,999` 0.146
`ppagecat_25 - 34` 0.120
`I40_$150,000 to $199,999` 0.111
race_5cat_White 0.108
female1 0.097
`ppagecat_55 - 64` 0.086
L0_a1 0.078
`ED0_Less than HS` 0.078
`ppagecat_35 - 44` 0.073
ED0_Professional 0.067
marrfemale1 0.067
singlefemale1 0.064
ppmarit_5 0.064
-2
-1
0
1
2
SHAP value (impact on model output)
Feature value
Low
High

Figure 5