Why No Booster?

Finding predictors to determine why people do not get the booster

By: Bilal Mukhtar

Goal

- To discover predictors that help understand why those who received the vaccine chose not to get the booster

Context

- As of right now, boosters are widely available for adults throughout the country

- However, many still choose not to receive these boosters, even if they chose not to get the vaccine

Process

- I looked at the 3 main reasons why people do not get the booster

- Then I tried to find the major predictors that lead to these reasons

Data

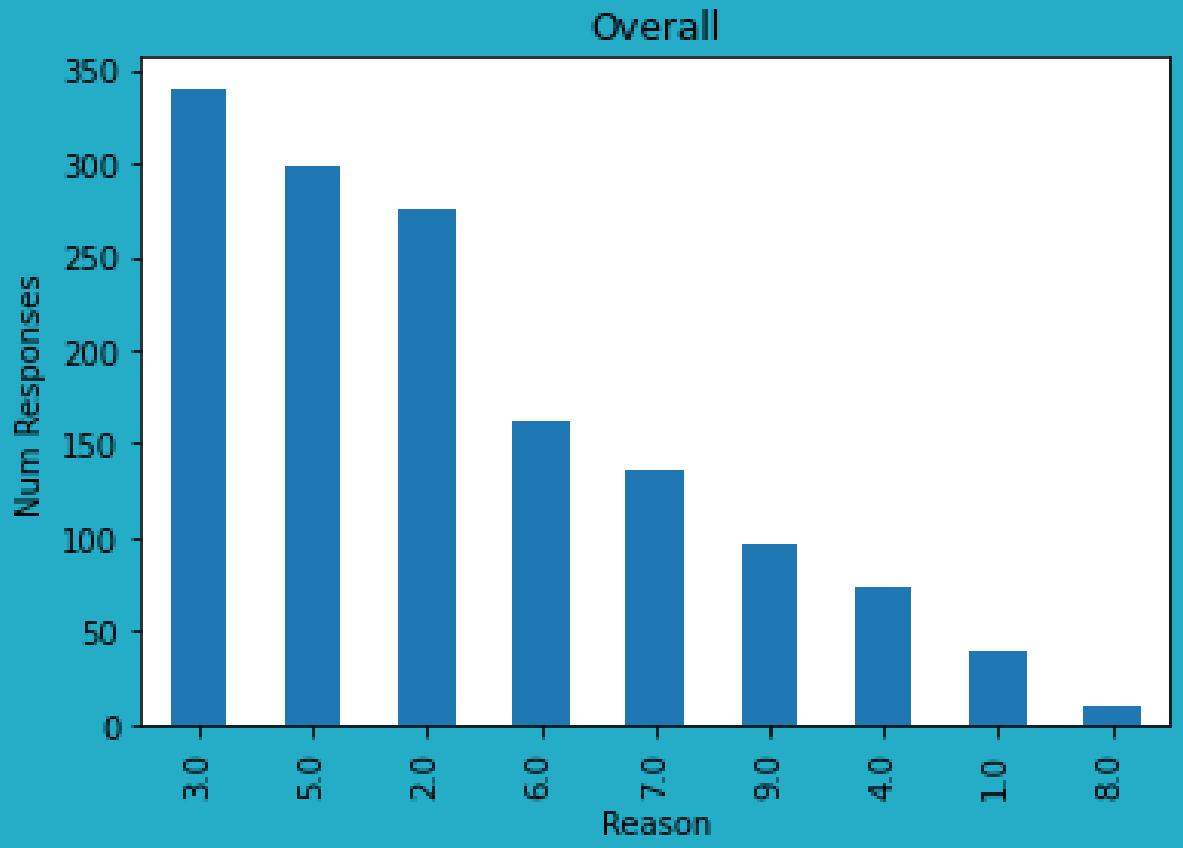
- All data was sourced from the Household Pulse Survey (HPS) Week 46 release

Goal Formation

- The goal came about from analyzing previous visualizations that showed the major reasons people chose not to get the booster

Previous Visualization

- The following graph shows the most common reasons for not getting the booster



Reasons Graph

Reasons for no Booster

- 2: Plan to get booster
- 3: Booster is unnecessary
- 5: Already had COVID

Data Cleaning

- All variables that were continuous values were removed
- Then any variable with over 80% null values were removed
- Lastly, NA values were imputed using the mean

Variable Selection

- Forward selection was used for variable selection
- Models were created for Reasons 2 and 3
- No model was created for Reason 5

Reason 5 Correlation

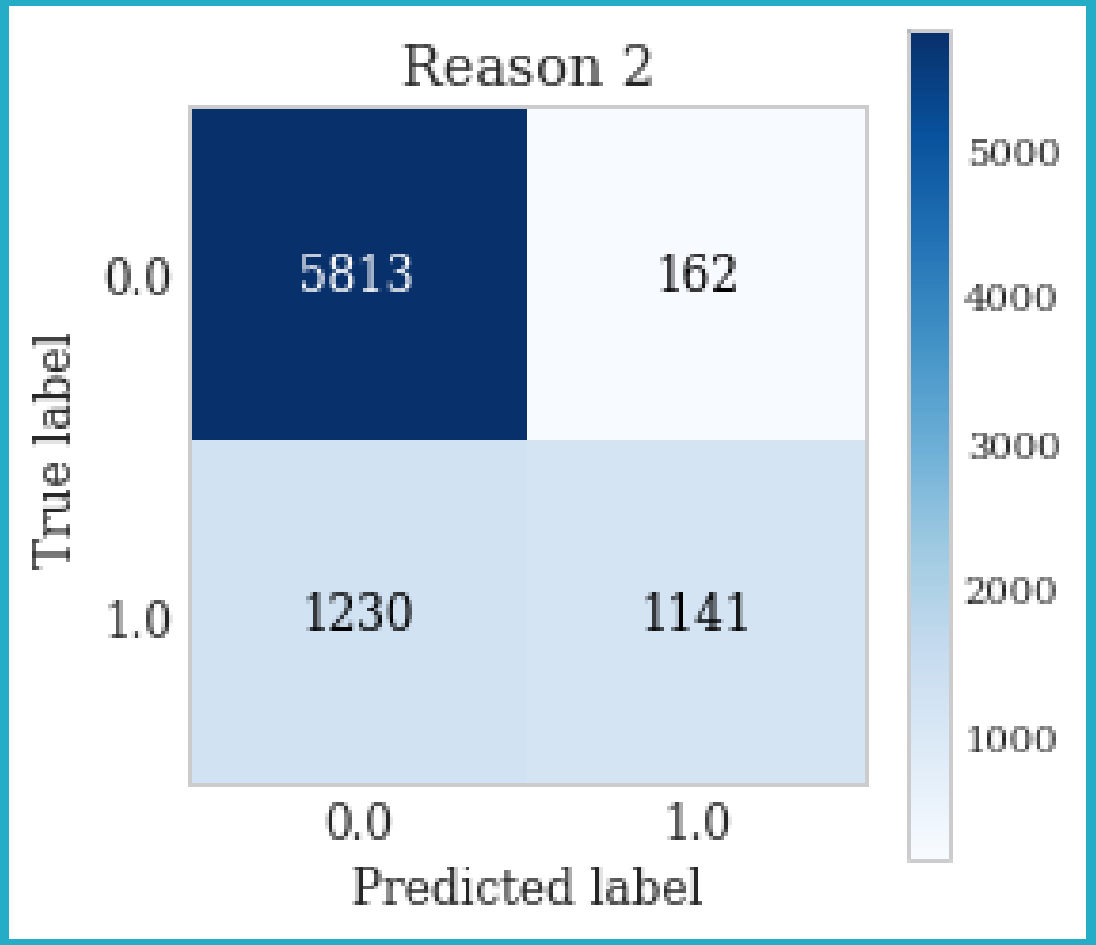
- Due to reason 5 being “Already had COVID”, any variable that also required having COVID was automatically a good predictor

- As such, these variables were removed from the set of possible predictors before variable selection

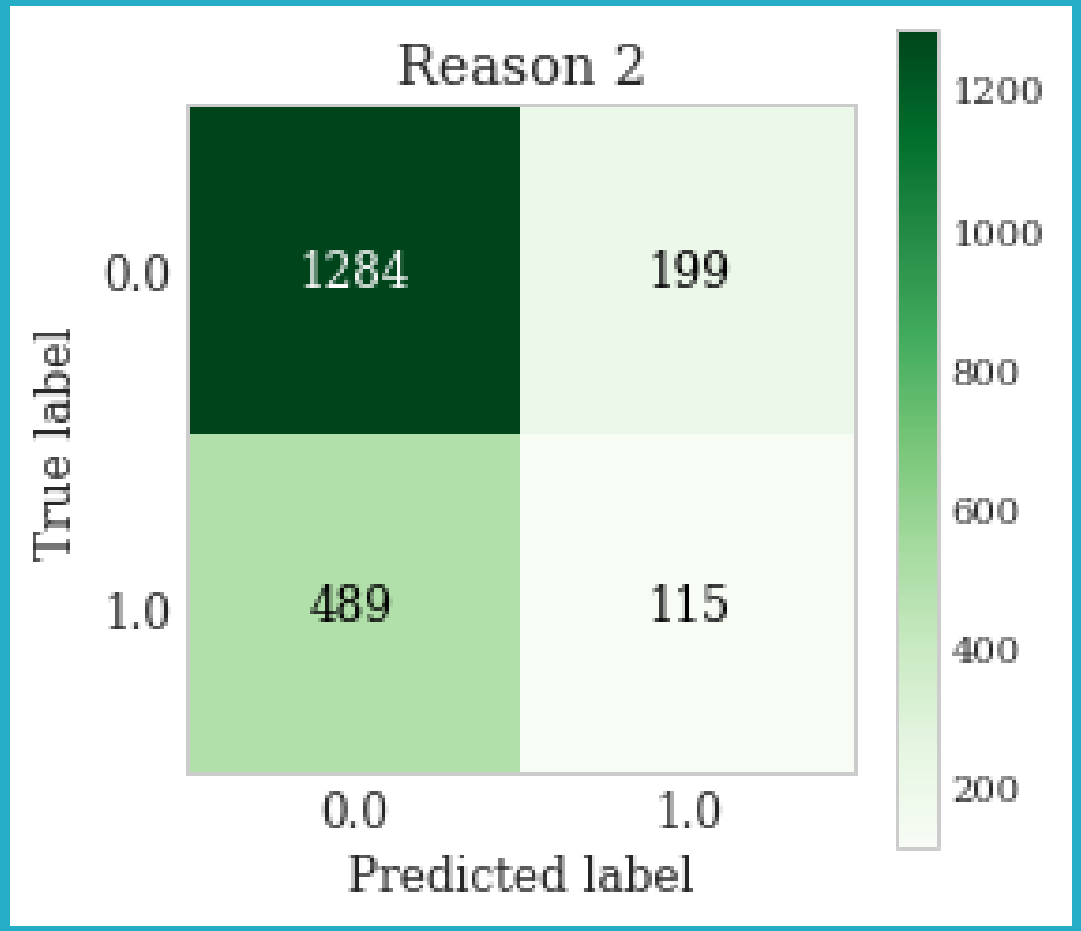
Modeling

- Each reason was split up for modeling with separate models being created due to separate variables being used
- Training and testing data sets were created, then confusion matrices and importance graphs were created from the models
- Bagging was used for making each model

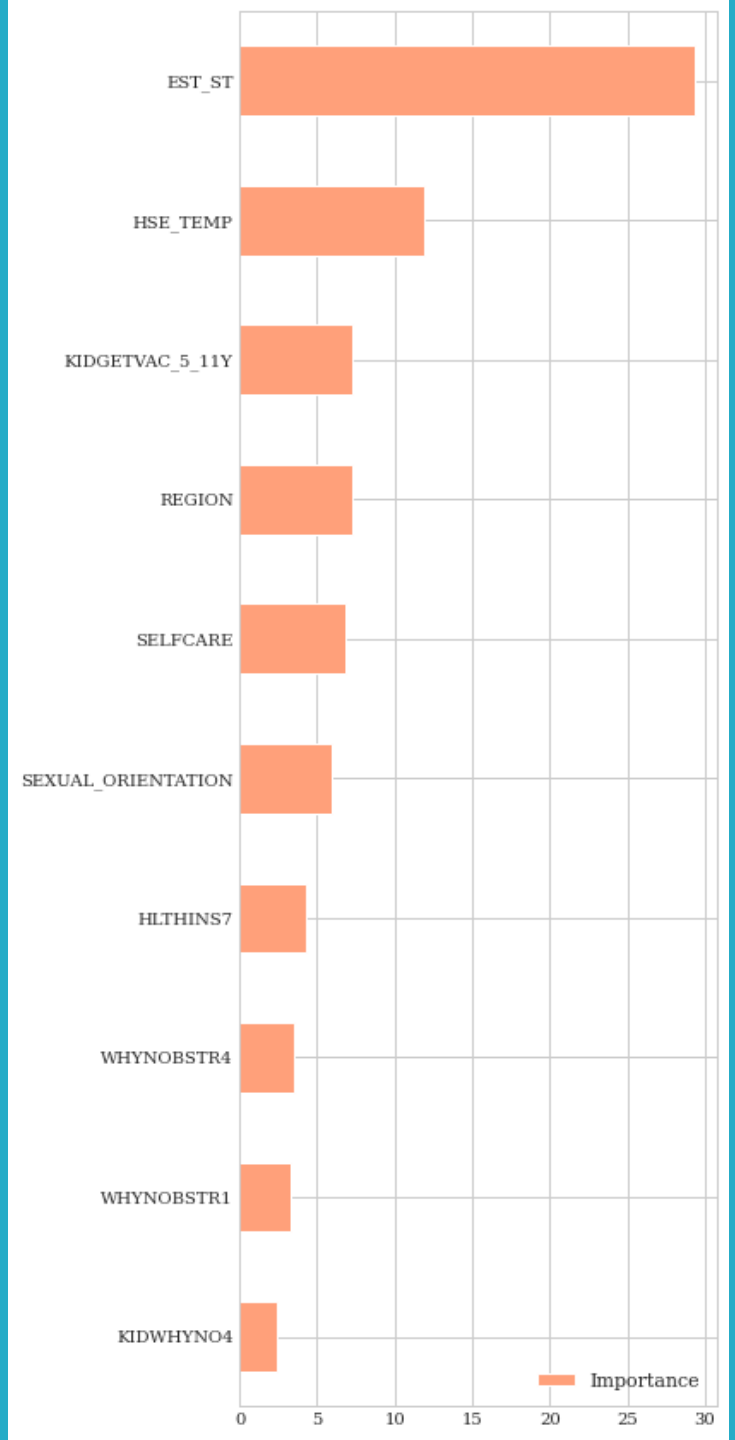
Reason 2 / Plan to get booster



Training Confusion Matrix



Testing Confusion Matrix



Variable Importance Graph

Stats

- Accuracy for the model was 83% on training data and 67% on test data

- EST_ST was clearly the most import variable and HSE_TEMP was a far second, with all other variables being unimportant

Variable Importance

- EST_ST: The state someone is from
- HSE_TEMP: Household was kept at an unsafe temperature due to cost

Variable Interpretation

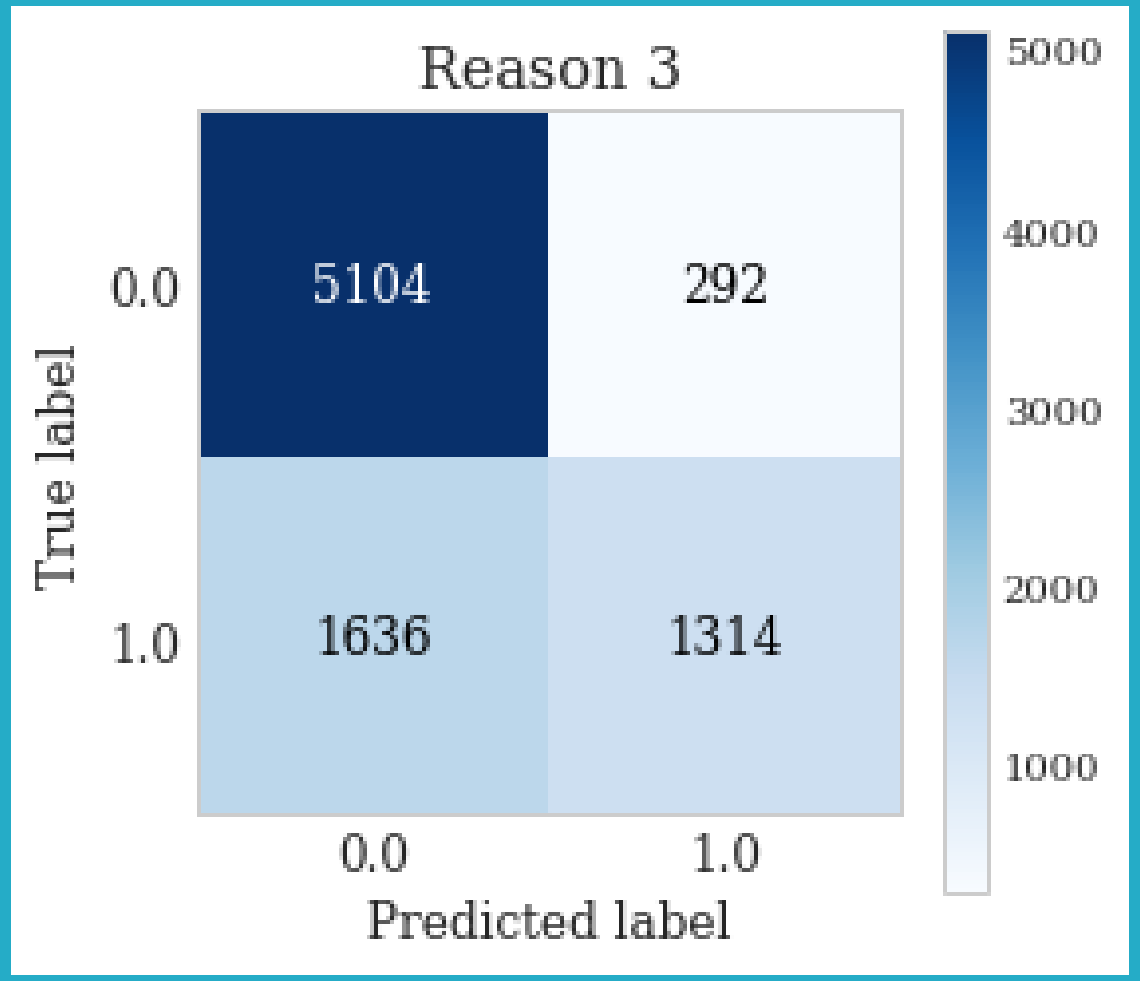
- The biggest determinate for someone planning to get the booster appears to be state
- HSE_TEMP probably tells very little, except that these states may be poorer states

EST_ST analysis

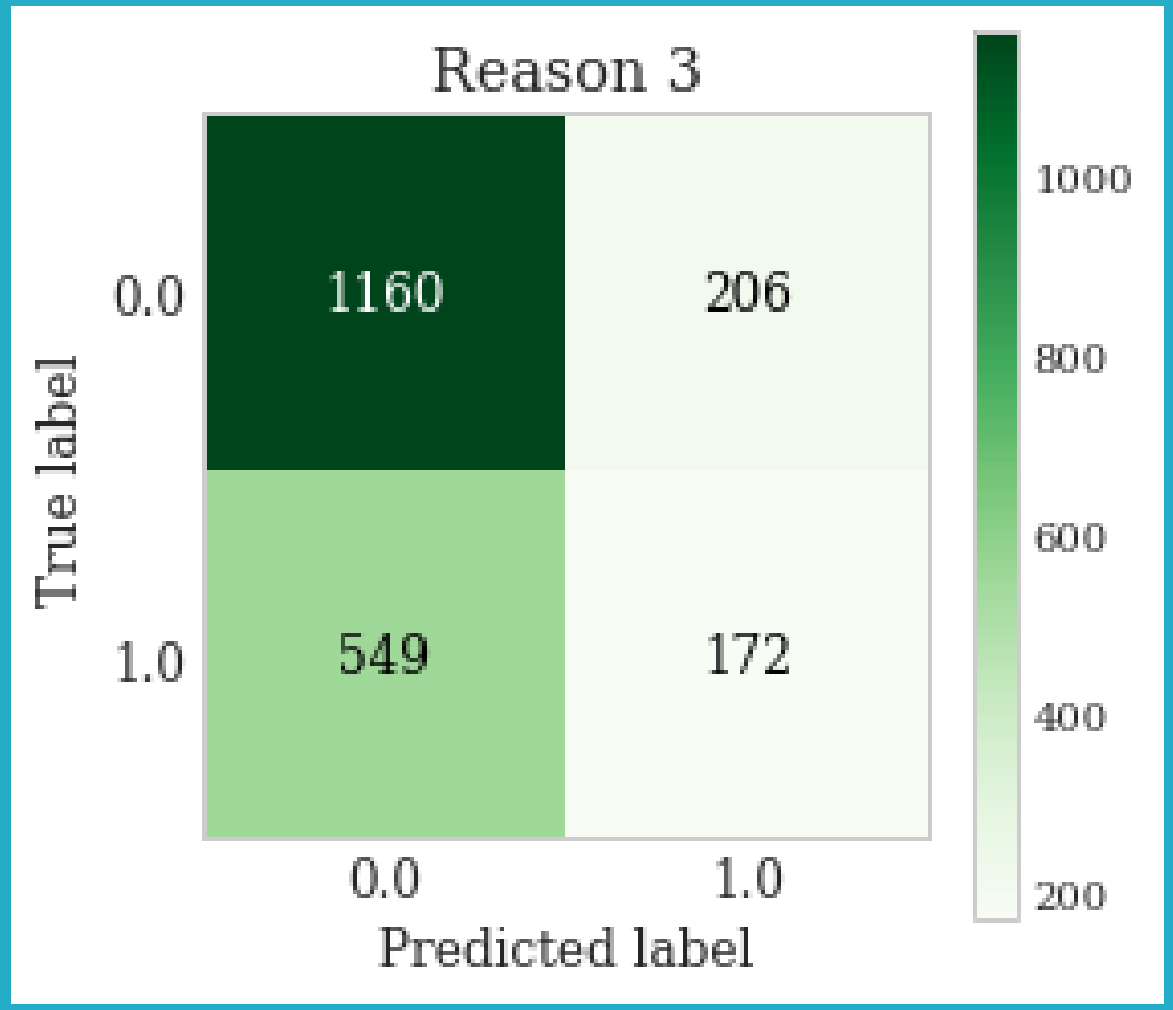
- After looking at the most common states, the top 3 are Texas, California, and Florida

- Being the 3 most populous states, it appears there is very little actual information that can be gathered from using the state someone is from

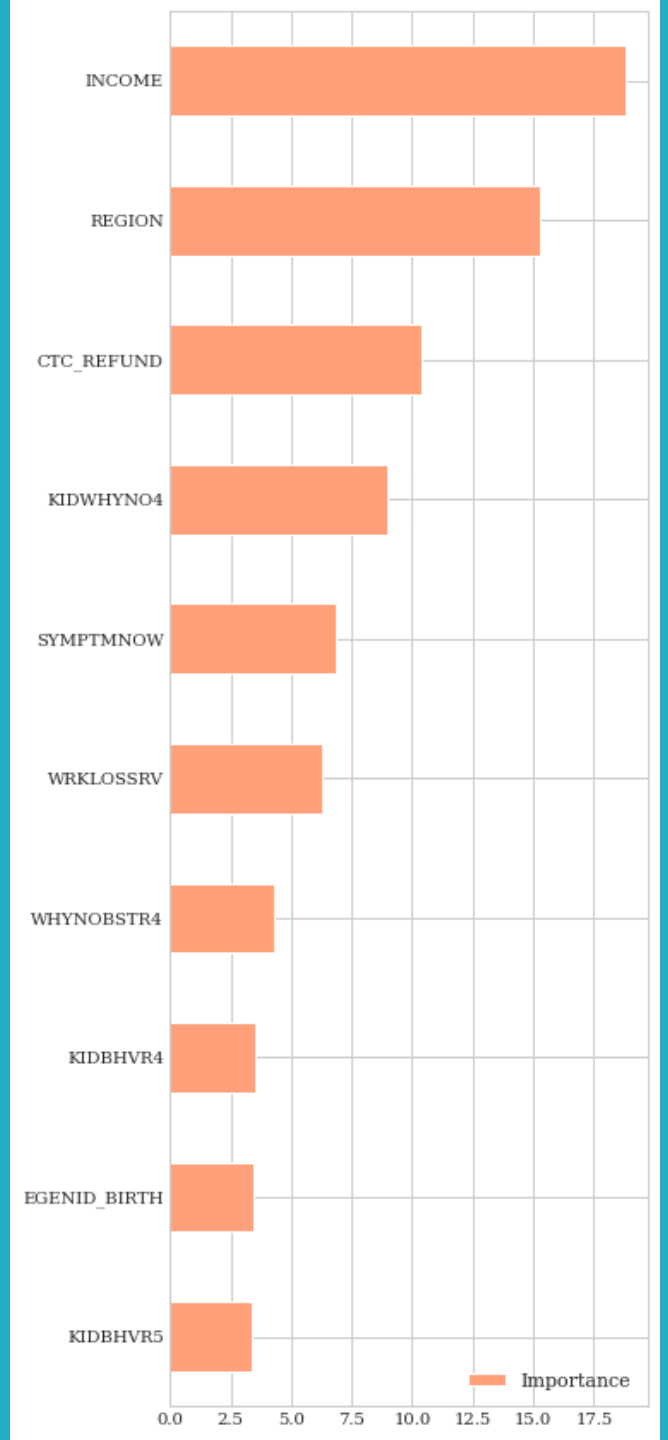
Reason 3 / Booster is unnecessary



Training Confusion Matrix



Testing Confusion Matrix



Variable Importance Graph

Stats

- Accuracy for the model was 77% on training data and 64% on test data

- INCOME, REGION, CTC_REFUND, and KIDWHYNO4 appear to be the most important variables, but there are multiple other variables with some level of importance

Variable Importance

- INCOME: Income in the past year
- REGION: Northeast, South, Midwest, or West of US
- CTC_REFUND: Received a tax refund
- KIDWHYNO4: Do not believe children need COVID vaccine

Variable Interpretation

- Using INCOME and REGION, these are probably people from a specific region and specific income
- CTC_REFUND probably has very little actual impact
- KIDWHYNO4 is the most obvious candidate, as someone who does not want to get the booster probably would not want their kid getting the vaccine

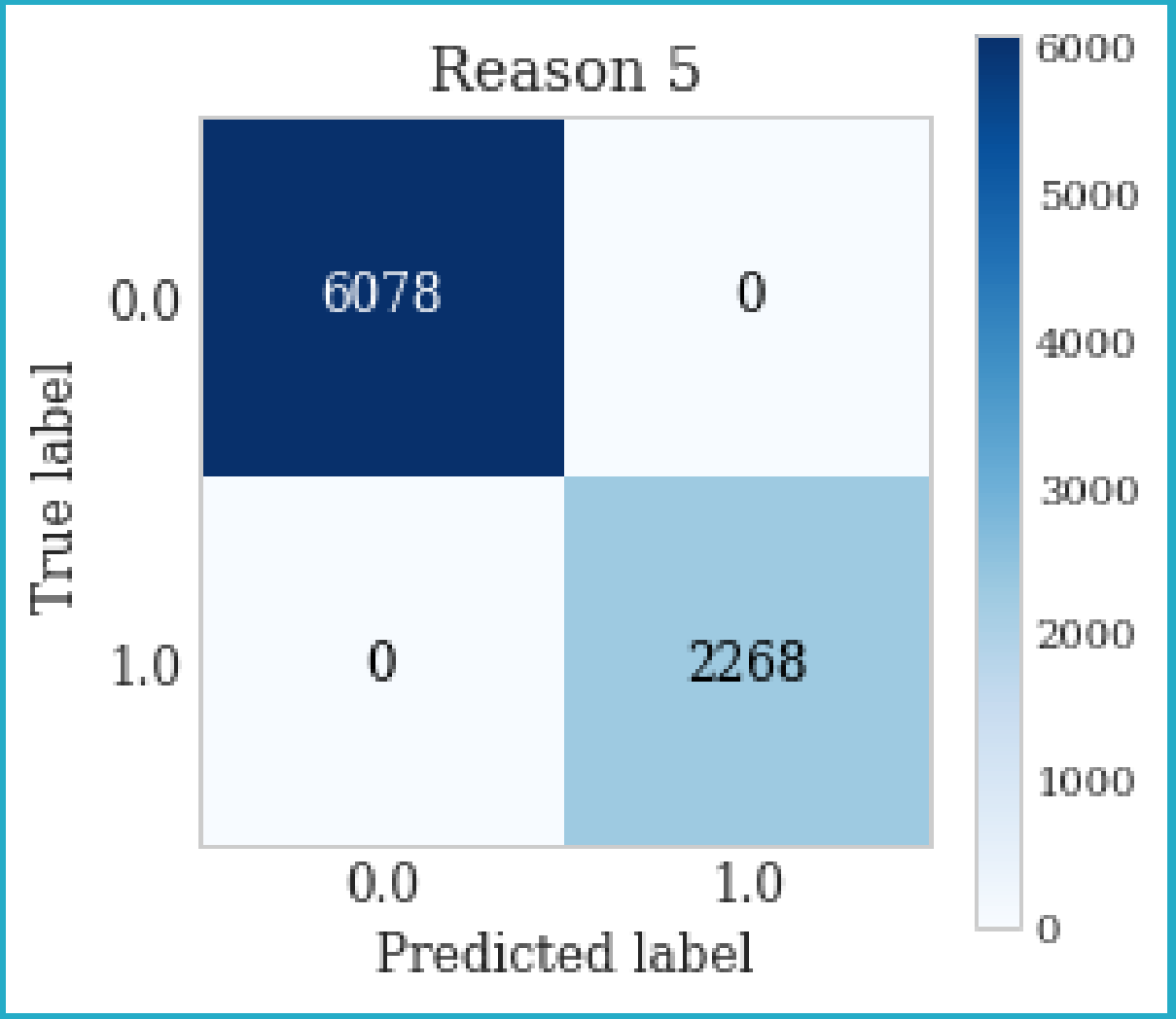
Income/Region analysis

- In terms of income, the most common were between $50,000 to $150,000 which shows about middle class
- The region was more telling, with most people being from the South, which is a region that is known to be anti-vaccine/booster

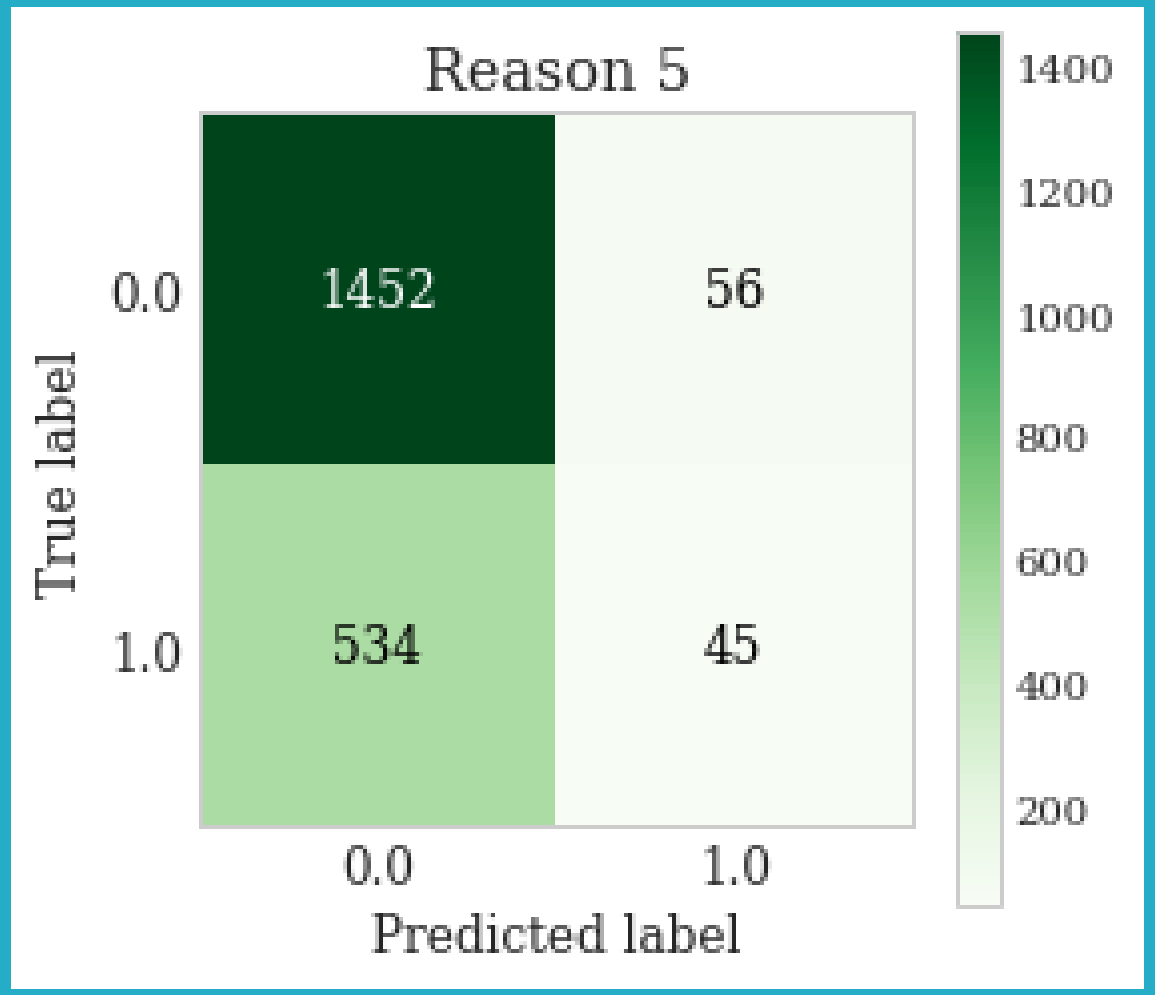
Reason 5 / Already had COVID

Training Data Issue

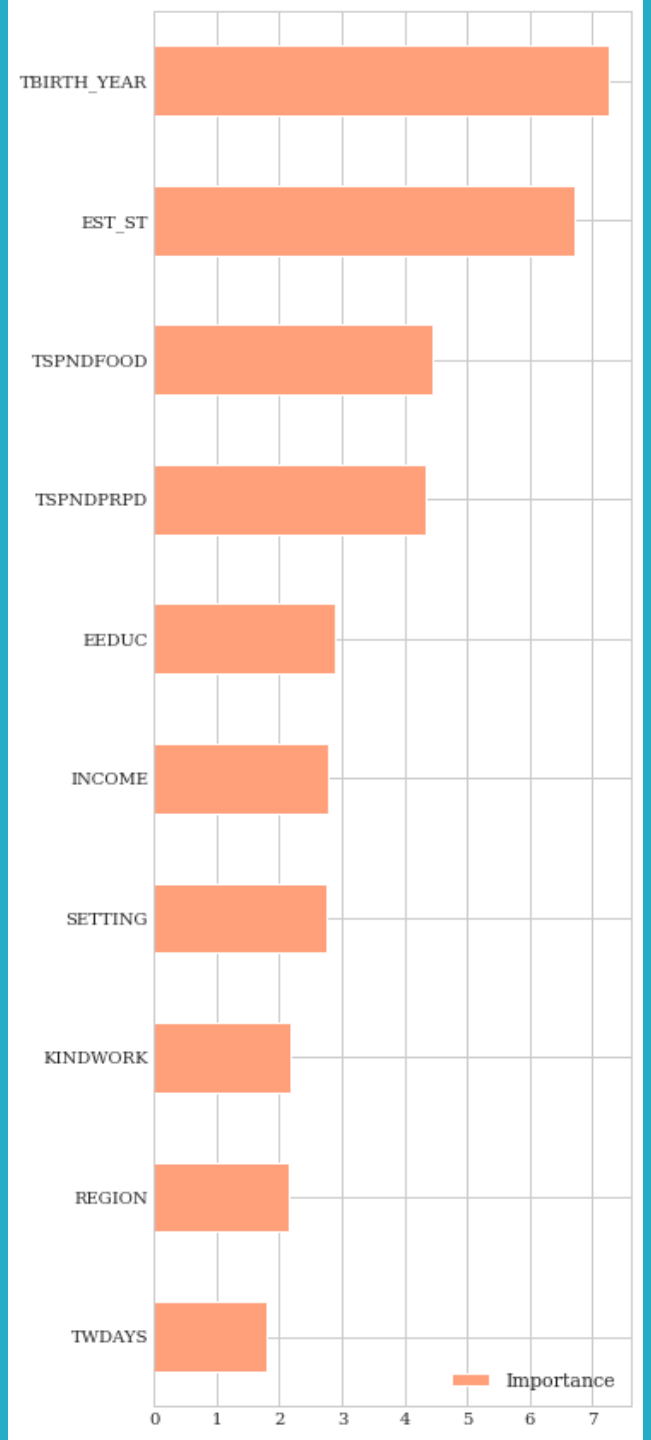
- Although I thought I removed any variables that were interacting, the training model still had 100% accuracy



Training Confusion Matrix



Testing Confusion Matrix



Variable Importance Graph

Stats

- Accuracy for the model was 100%* on training data and 72% on test data

- TBIRTHYEAR and EST_ST were the only predictors with any importance

Variable Importance

- EST_ST: State person is from

- TBIRTHYEAR: Year person was born

Variable Interpretation

- Due to the overall low importance of the variables for this model, neither TBIRTHYEAR or EST_ST appear to show anything important

EST_ST/TBIRTH_YEAR analysis

- EST_ST again just showed the most populous states

- TBIRTH_YEAR showed the around age 40 in terms of year, which is a common age for those responding to the survey

Conclusions

- Reasons 2 and 5 appear to have no significant variable to predict them

- Reason 3, however, appears to have a couple variables that would also be indicative of someone from a conservative region/state

Future Application

- Mainly for reason 3, the important predictors could be used to target campaigns to promote the booster

- This could be done for the other reasons, although this may be ineffective