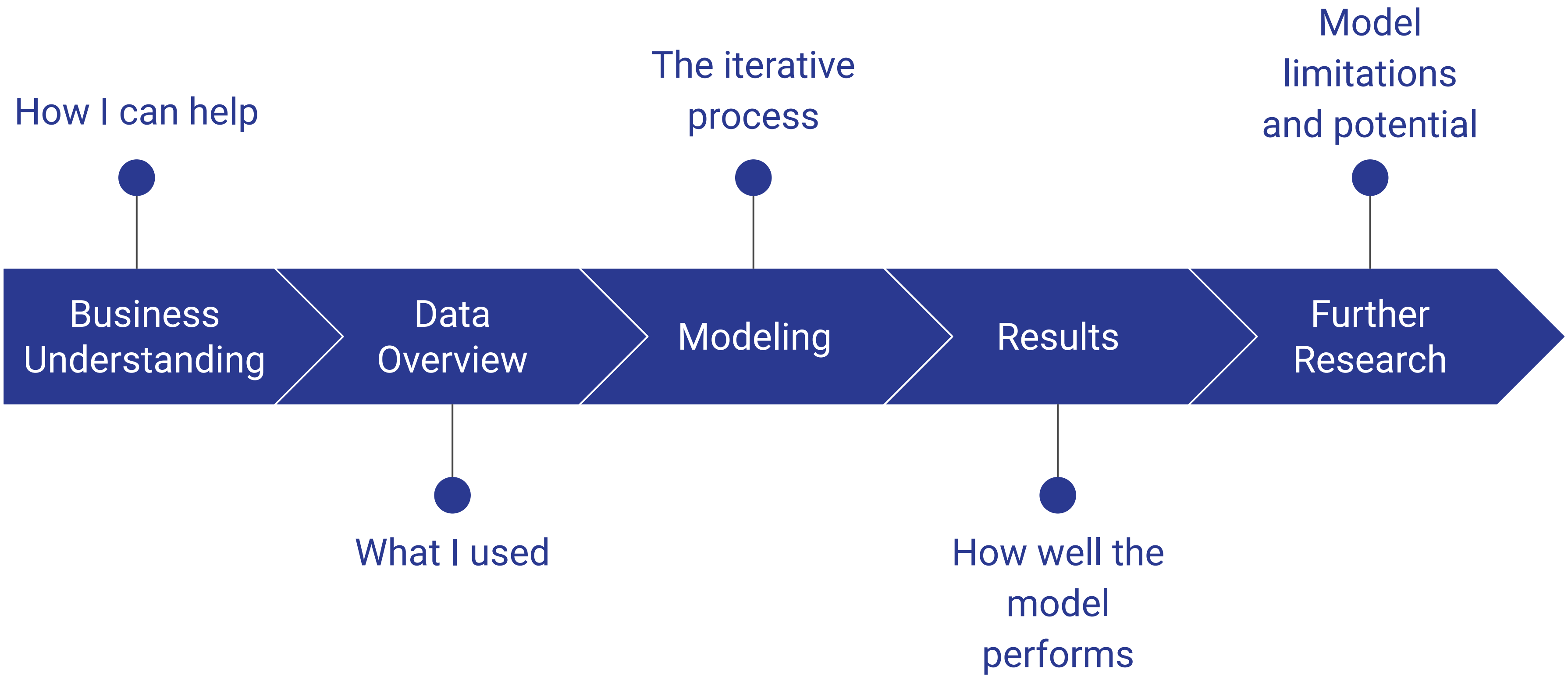
Who Wants Shots?

Machine Learning to Predict Vaccination Status

Phase 3 Project
By Ashley Eakland

Agenda



Business Understanding

How I can help

Data Overview

What I used

Modeling

The iterative process

Results

How well the model performs

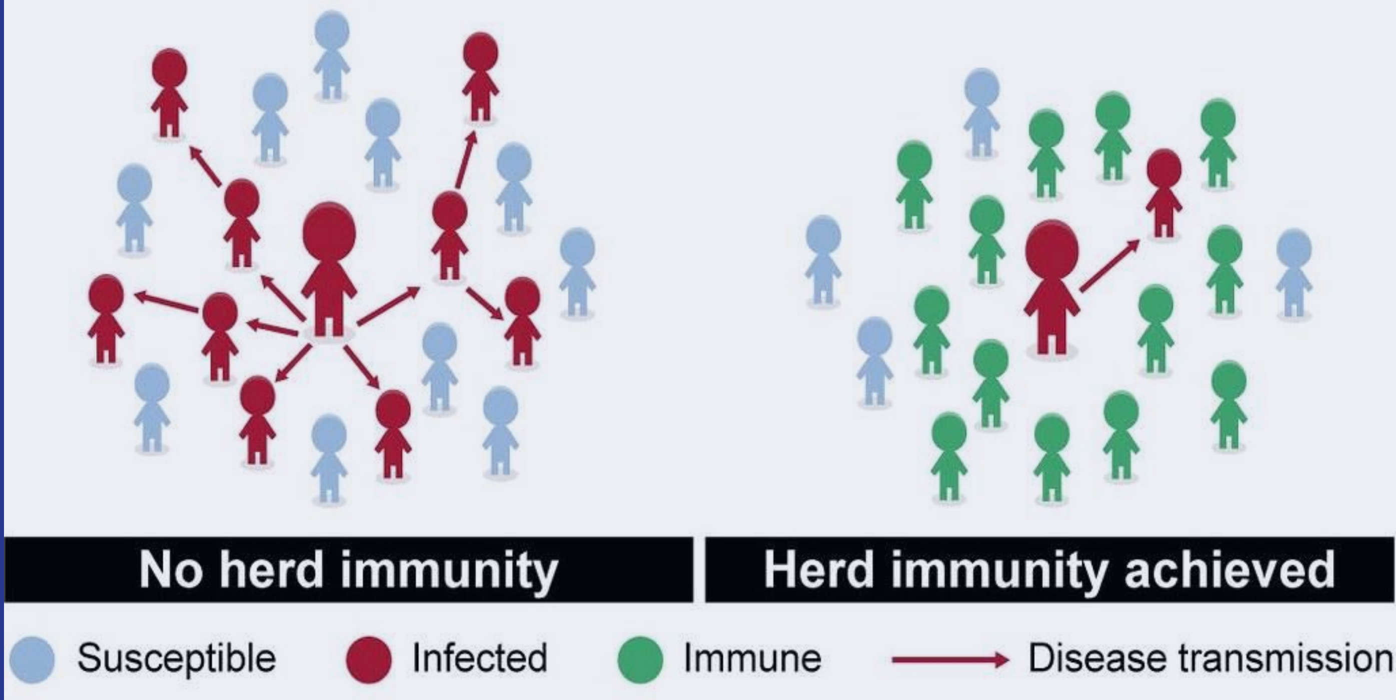
Further Research

Model limitations and potential

Business Problem

Guiding public health efforts with regard to vaccination status utilizing predictive modeling

Why?



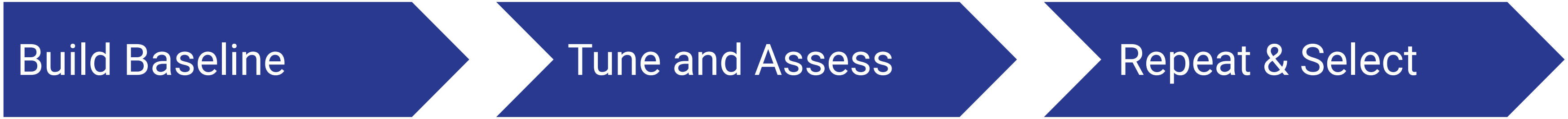
Source: GAO adaptation of NIH graphic. | GAO-20-646SP

Data Understanding

Data provided by the US National Center for Health Statistics

- National 2009 H1N1 Flu Survey
  - 37 question survey
  - 26,700+ respondents
  - 6,500 complete surveys
- Seasonal Flu
- Yes/No and small scale rankings
- Encoded values for anonymity

Modeling Process



Build Baseline

Tune and Assess

Repeat & Select

Base

- Build a base model for comparison

Iterations

- Parameter Tuning
    - Assess metrics

Final Model

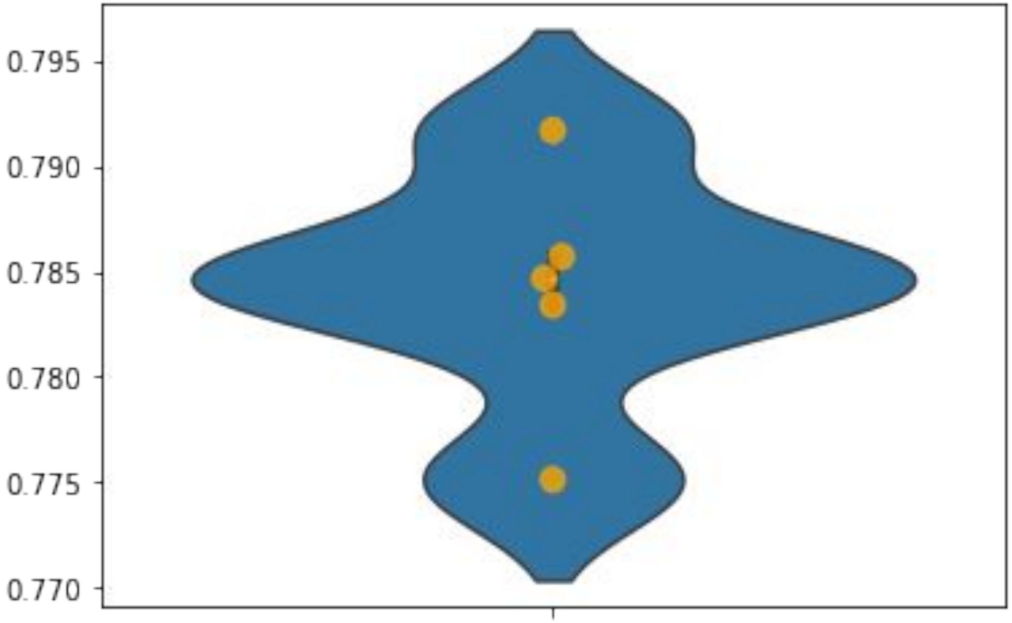
- Model with optimal performance is selected

Modeling Process *cont.*

Build Baseline

Base - RandomForest

- Distribution of validation scores

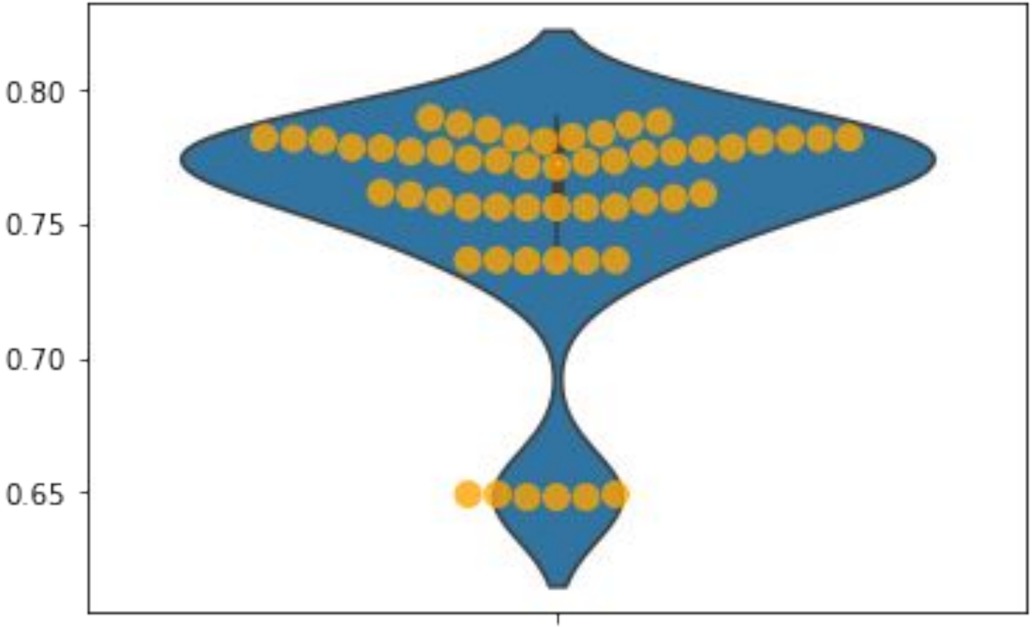


Modeling Process

Tune and Assess

RandomForest Iterations

- Tuned parameters and added folds for assessment

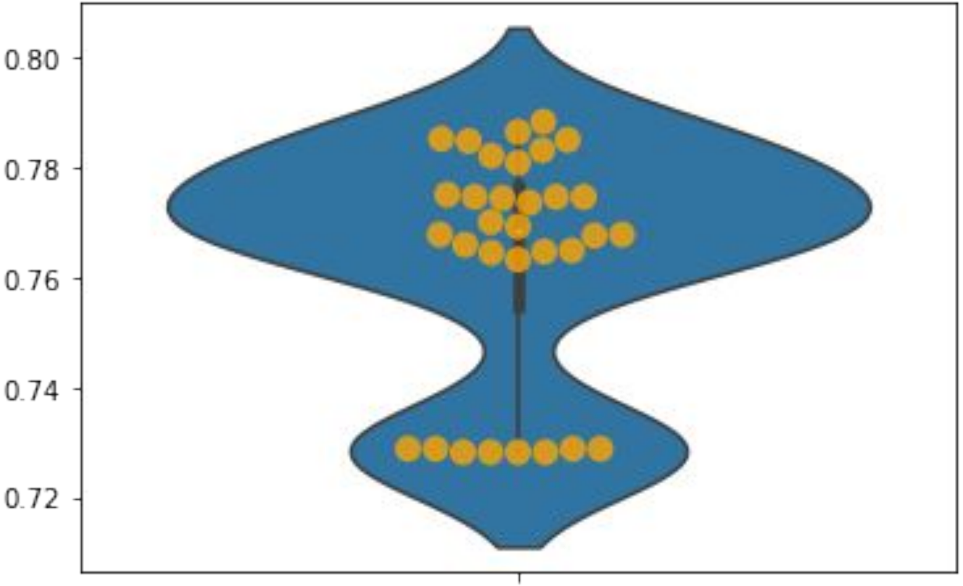


Modeling Process

Repeat & Select

RandomForest Final Model

- Model with optimal performance is selected based on target metrics

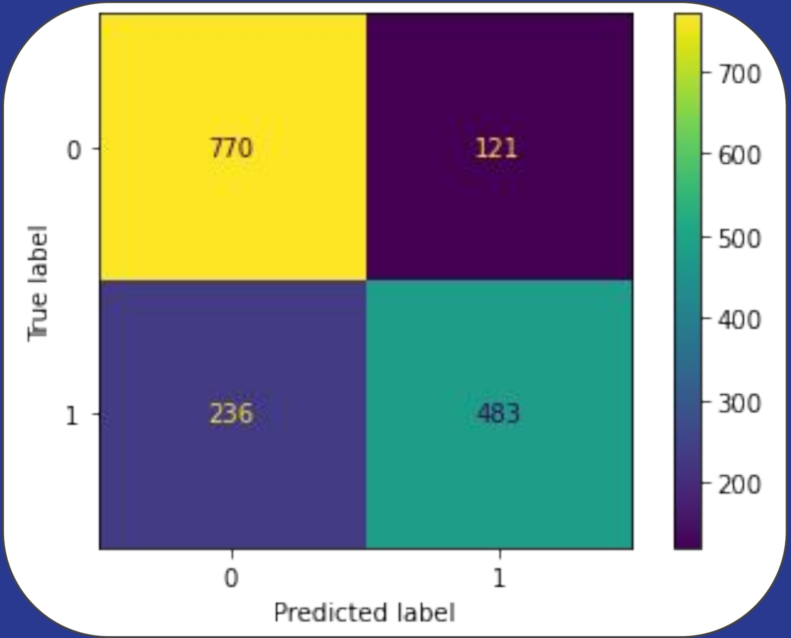


Final Results

RandomForest

78%

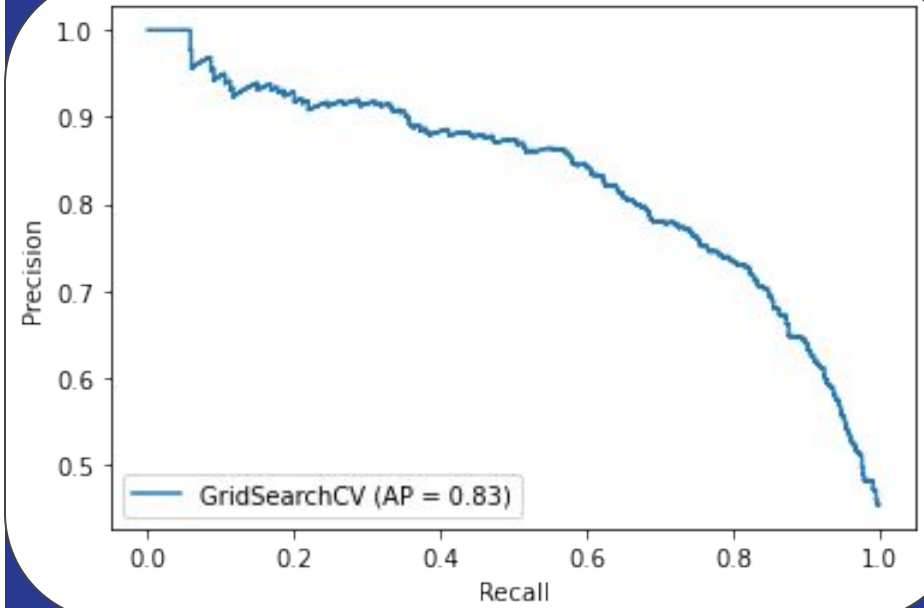
Accuracy



Precision

80% precise labeling

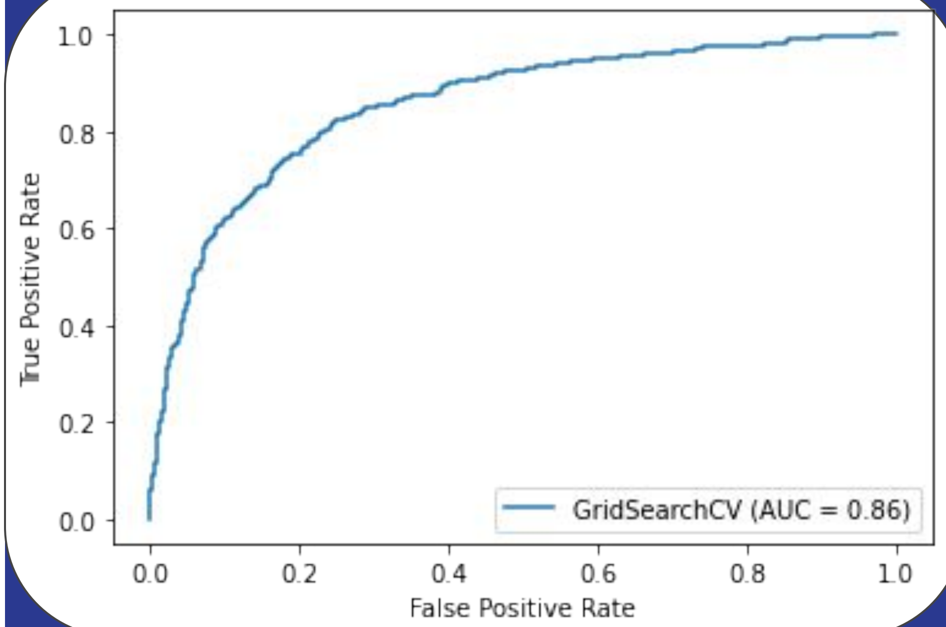
If model predicted a given sample as vaccinated, it’s 80% probability that the model is correct.



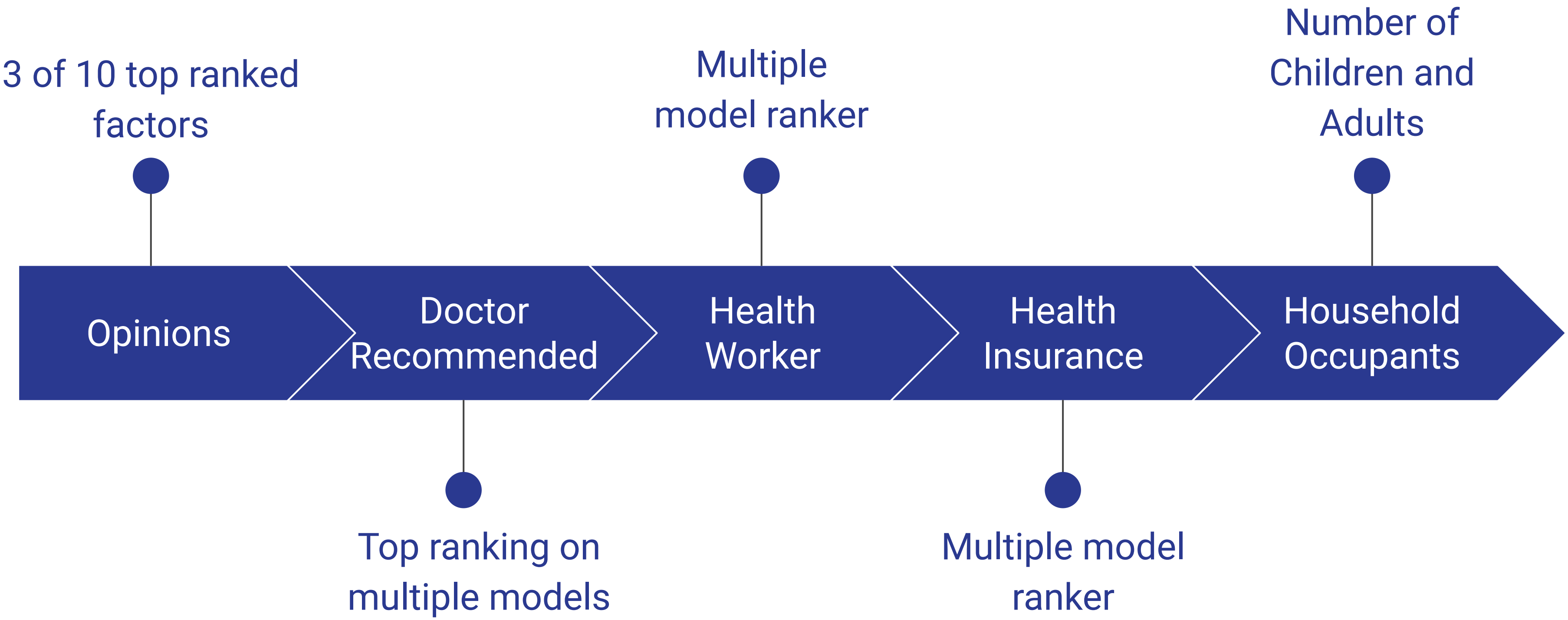
AUC

0.86

1 = perfect classification
0.50 = no better than random guessing



Top Factors



Opinions

3 of 10 top ranked factors

Doctor Recommended

Top ranking on multiple models

Health Worker

Multiple model ranker

Health Insurance

Multiple model ranker

Household Occupants

Number of Children and Adults

** Employment industry also appeared at the bottom of the top ten - to be discussed with further analysis*

Future Analysis

Limitations and future improvement steps

Limitations

- **Not for predicting COVID-19 data**

- Employment Occupation & Industry worth further exploration - both have at least 1 appear on top 10 important features

Future Analysis *cont.*

All model types perform consistently, varying less than 1-3% on target metrics.

Future Improvements

- Include parameter to allow unknown encoded employment categories
- Refine structure of features for cleaner binary classification
- Refactor with incomplete survey responses - binned “prefer not to answer”

Future Analysis *cont.*

H1N1 Survey Responses were also analyzed and models prepared

❏ Data had severe class imbalance

❏ Model Performance less consistent, though could be improved with more time

H1N1 Specific Future Improvements

- Refine structure of features for cleaner binary classification

- Reattempt with a pared down dataset of less incomplete survey responses

Thank you!

Ashley Eakland | ashley@eakland.net

LinkedIn | *https://www.linkedin.com/in/ashleyeakland/*

For technical information and to see the Jupyter Notebooks:

GitHub Repository | *https://github.com/smashley-eakland/who-wants-shots*