Predicting H1N1 and Seasonal Flu Vaccination

Lavanya Acharya
Katrin Ayrapetov
Sean Li

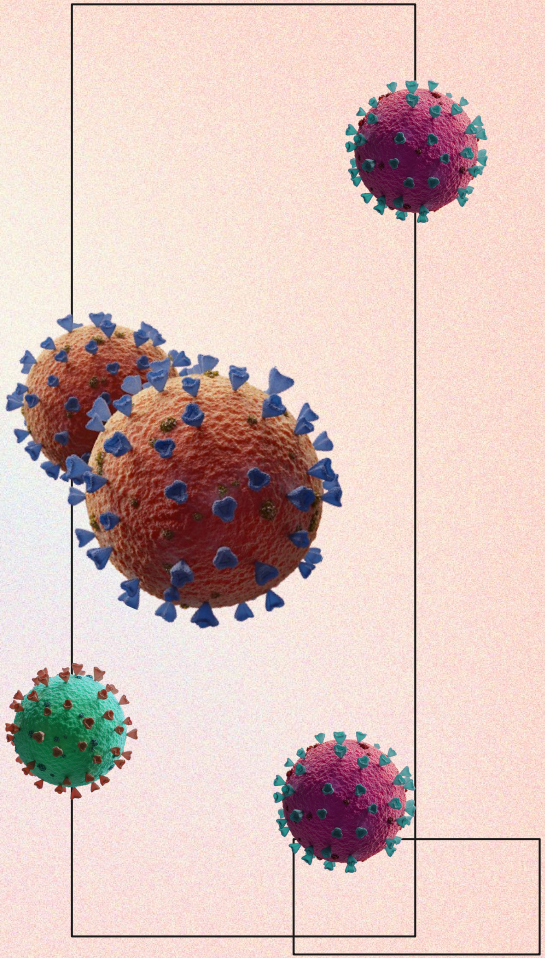
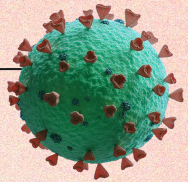


TABLE OF CONTENTS



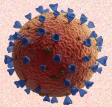
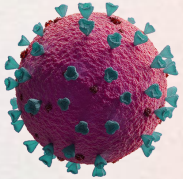


TABLE OF CONTENTS

01 BACKGROUND



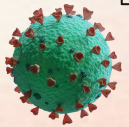
The 2009 flu pandemic in the United States was caused by a novel strain of the Influenza H1N1 virus, commonly referred to as “swine flu.”

From April 2009 to April 2010, the CDC estimates in the United States there were

- 60.8 million cases
- 274,304 hospitalizations
- and 12,469 deaths

A vaccine for the H1N1 flu virus became publicly available in October 2009.

A view of a newspaper headline near Times Square in New York, New York, USA, on 27 April 2009. Photo by EPA/BGNES



01 BACKGROUND

An H1N1 flu vaccination clinic held in San Francisco in December 2009, Justin Sullivan/Getty Images

From August 2009 to May 2010, one or more doses of the H1N1 vaccine were administered to

- 29.1 million children
- 80.8 million adults

From August 2009 to May 2010, one or more doses of the seasonal flu vaccine were administered to

- 31.6 million children
- 91.6 million adults

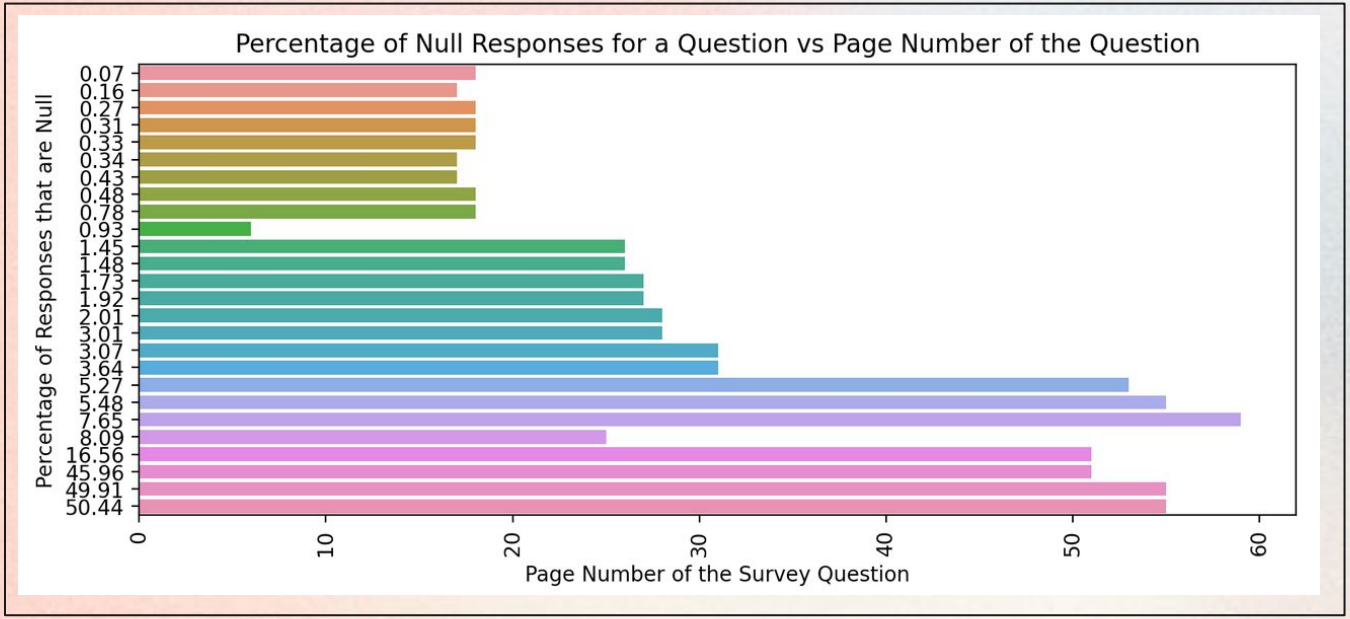
01 BACKGROUND

The US National Center for Health Statistics conducted National H1N1 Flu Survey.

The survey was conducted by phone and the survey takers asked respondents whether they had received the H1N1 and seasonal flu vaccines, in conjunction with questions about themselves.

The questions covered the respondents' social, economic, and demographic background, opinions on vaccine effectiveness and risks, and behaviors towards mitigating transmission.

01 BACKGROUND

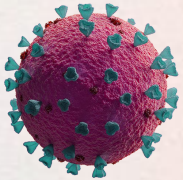


The data had null values in two cases:

- Respondent choosing "no response" or "do not know" as a response
- Respondent not answering the question

The survey is sixty pages long. The later a particular question appeared in the survey, the more null values there were for the responses.

01 PROBLEM STATEMENT



Goals:

1. Predict whether respondents received the H1N1 and seasonal flu vaccines, using the US National Center for Health Statistics survey data.
2. Identify the most predictive features for choosing not to vaccinate.
3. Using these, create an abridged version of the survey that has the same predictive power as the original full size survey.

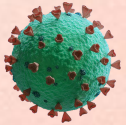
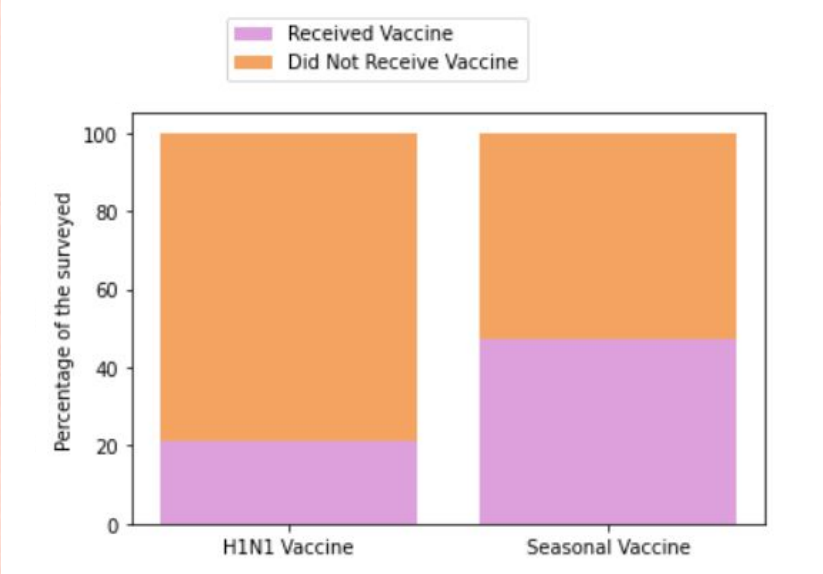


TABLE OF CONTENTS

02 SUMMARY OF TARGET VARIABLES

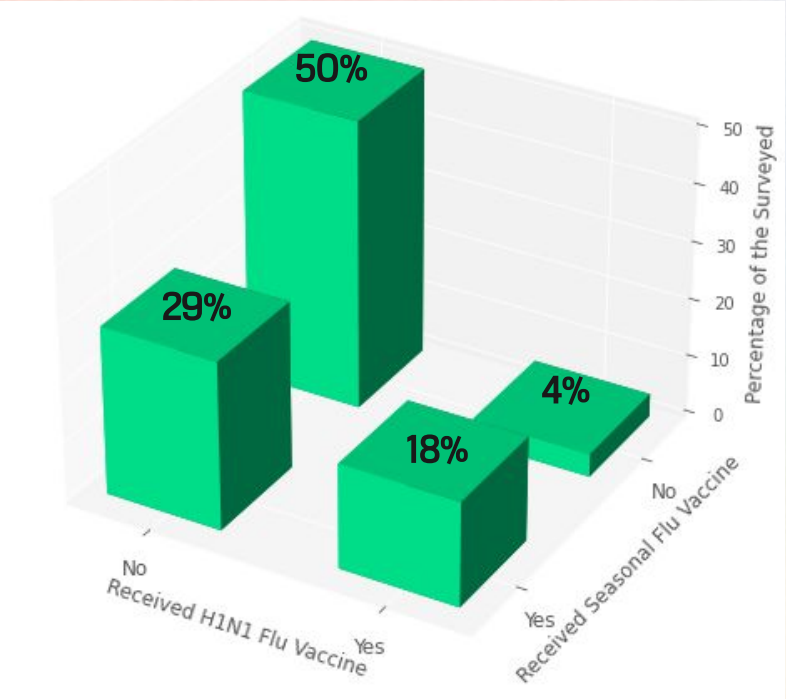


In the training data set:

- 21% **received** the H1N1 vaccine
- 79% **did not receive** the H1N1 vaccine

- 47% **received** the seasonal vaccine
- 53% **did not receive** the seasonal vaccine

02 SUMMARY OF TARGET VARIABLES



The two target variables have been shown to have statistically significant dependance.

The probability that a randomly selected respondent:

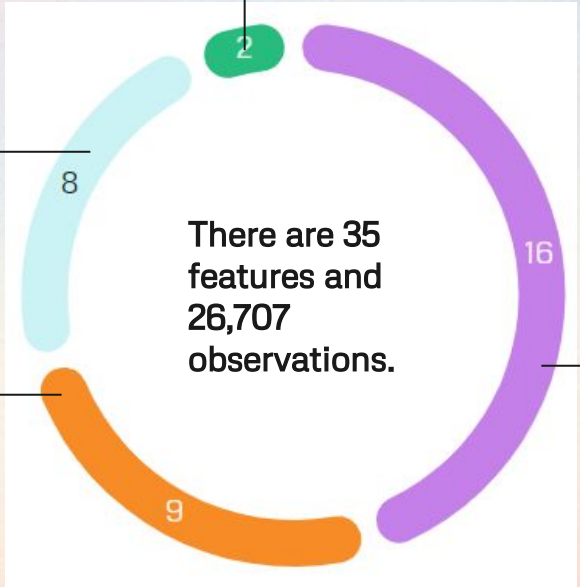
- Received the seasonal vaccine is 47%
- Received the seasonal vaccine given they received the H1N1 vaccine is 82%

02 OVERVIEW OF FEATURE VARIABLES

Numeric features are the number of adults and number of kids in the survey taker's household

Ordinal features which give the survey taker's opinions like their feelings about the vaccine efficacy and risk.

Nulls were imputed with the median of the responses.



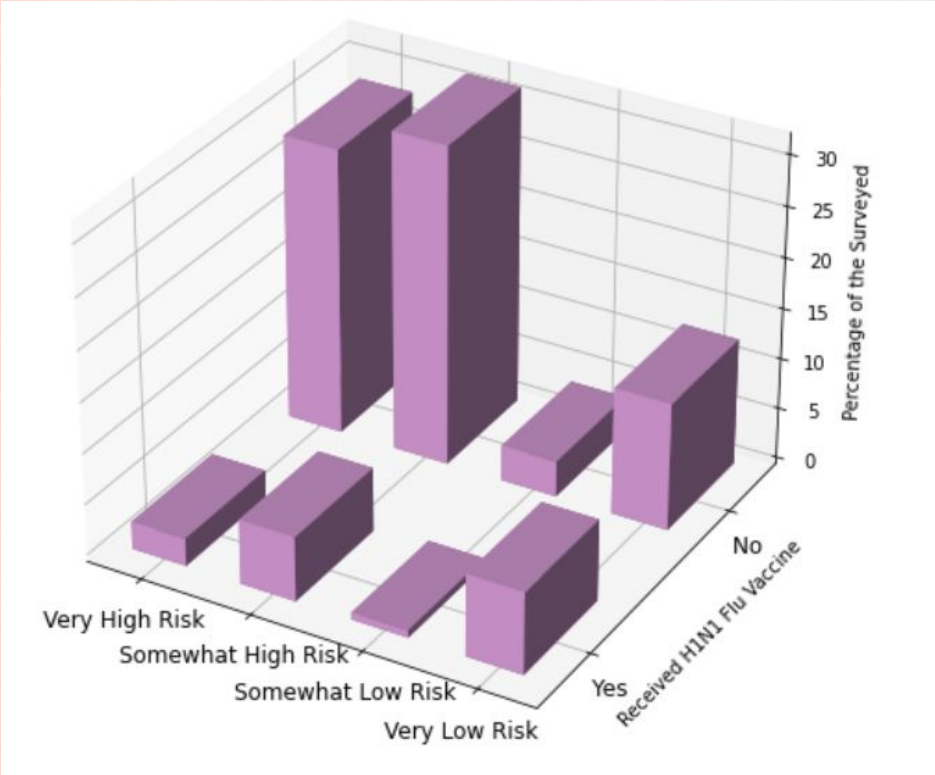
Categorical features which give demographic information about like the survey taker's age, employment status and income.

Nulls were imputed with "no response".

Categorical features which give information about the survey taker's behaviors like hand washing and mask wearing.

Nulls were imputed with "no response".

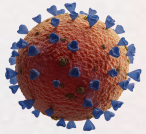
02 OVERVIEW OF FEATURE VARIABLES



There is statistically significant dependence between **respondents' perceived risk of the H1N1 vaccine** and whether or not they chose to receive it.

TABLE OF CONTENTS

03 OVERVIEW OF APPROACH

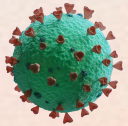


ID	Age	Marital Status	...	H1N1 Vaccine	Seasonal Flu Vaccine
0	55 - 64 Years	Not Married	...	yes	no
1	35 - 44 Years	Not Married	...	no	yes
2	18 - 34 Years	Not Married	...	no	yes
3	65+ Years	Not Married	...	no	yes
4	45 - 54 Years	Married	...	yes	yes
5	65+ Years	Married	...	no	no
...	...	...	...	...	...

We approached the multilabel classification problem by converting the data set into two single class binary datasets and fitting a binary classification model to each data set.

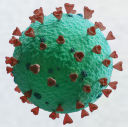
ID	Age	Marital Status	...	H1N1 Vaccine
0	55 - 64 Years	Not Married	...	yes
1	35 - 44 Years	Not Married	...	no
2	18 - 34 Years	Not Married	...	no
3	65+ Years	Not Married	...	no
4	45 - 54 Years	Married	...	yes
5	65+ Years	Married	...	no
...	...	...	...	...

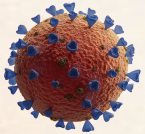
ID	Age	Marital Status	...	Seasonal Flu Vaccine
0	55 - 64 Years	Not Married	...	no
1	35 - 44 Years	Not Married	...	yes
2	18 - 34 Years	Not Married	...	yes
3	65+ Years	Not Married	...	yes
4	45 - 54 Years	Married	...	yes
5	65+ Years	Married	...	no
...	...	...	...	...



03 PRELIMINARY MODELING

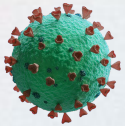
Using a gridsearch, we used all the features to see how a basic model would perform. Tests were ran for both seasonal flu and H1N1, using accuracy and AUC as metrics on logistic regression, *k*-NN, multinomial Naive Bayes, and random forest classifiers for a total of 16 tests.

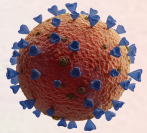




Logistic regression		
	Seasonal Flu	H1N1
Accuracy	0.779	0.848
AUC	0.854	0.856

03 USING SEASONAL FLU AS A PREDICTOR





Next we tried using seasonal flu vaccines as a predictor for H1N1, which offered a small but significant performance boost

	Preliminary Model	Using seasonal flu as a predictor
Logistic Regression	0.855	0.884
k-NN	0.798	0.820
mNB	0.789	0.803
Random Forest	0.854	0.886

Sadly, using predicted seasonal flu did not yield the same results, and the models performed similarly to the preliminary models

03 BACK TO THE DRAWING BOARD: METRICS

Because we are more interested in people *who did not get the vaccine*, we made them our positive class.

1: Did not receive the vaccine
0: Received the vaccine

Accuracy
The percentage of respondents classified correctly as having received the vaccine or not

Recall
The percentage of respondents who did not receive the vaccine who were classified correctly.

AUC
Measure of overall classification performance

Precision
The percentage of respondents were classified as not having received the vaccine who did not receive the vaccine.

03 INITIAL MODELS: H1N1 Vaccine

Logistic Regression Model

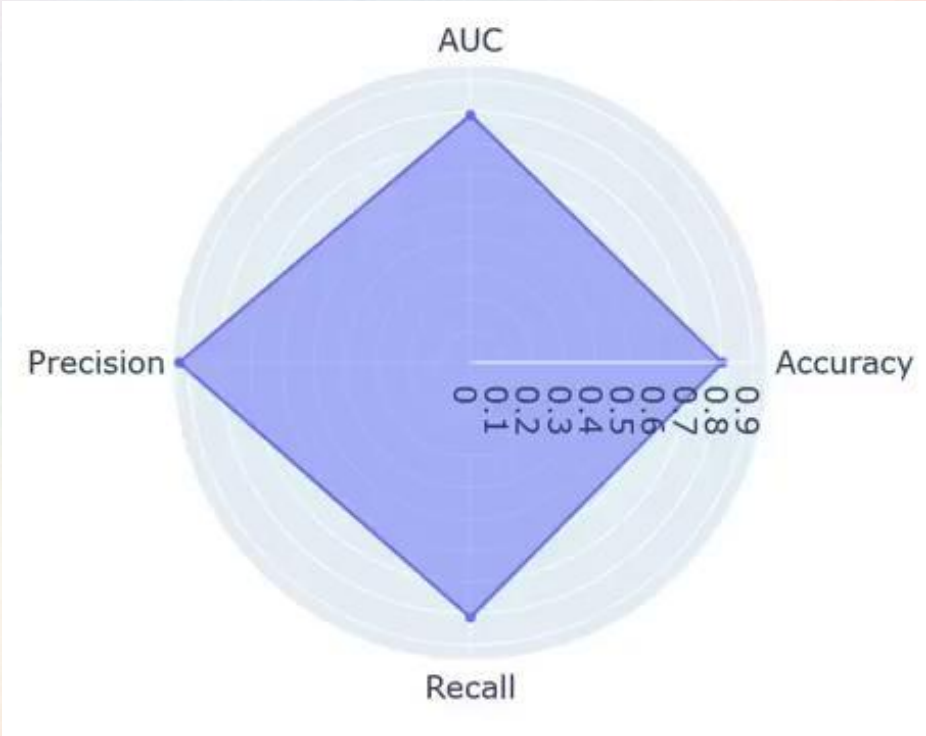
Accuracy: 0.803
AUC: 0.789
Precision: 0.928
Recall: 0.813

XGBoost Model

Accuracy: 0.8518
AUC: 0.730
Precision: 0.879
Recall: 0.942

Neural Network Model

Accuracy: 0.823
AUC: 0.897
Precision: 0.790
Recall: 0.873



03 INITIAL MODELS: Seasonal Vaccine

Logistic Regression Model

Accuracy: 0.789
AUC: 0.788
Precision: 0.804
Recall: 0.799

XGBoost Model

Accuracy: 0.797
AUC: 0.794
Precision: 0.800
Recall: 0.825

Neural Network Model

Accuracy: 0.800
AUC: 0.864
Precision: 0.800
Recall: 0.834

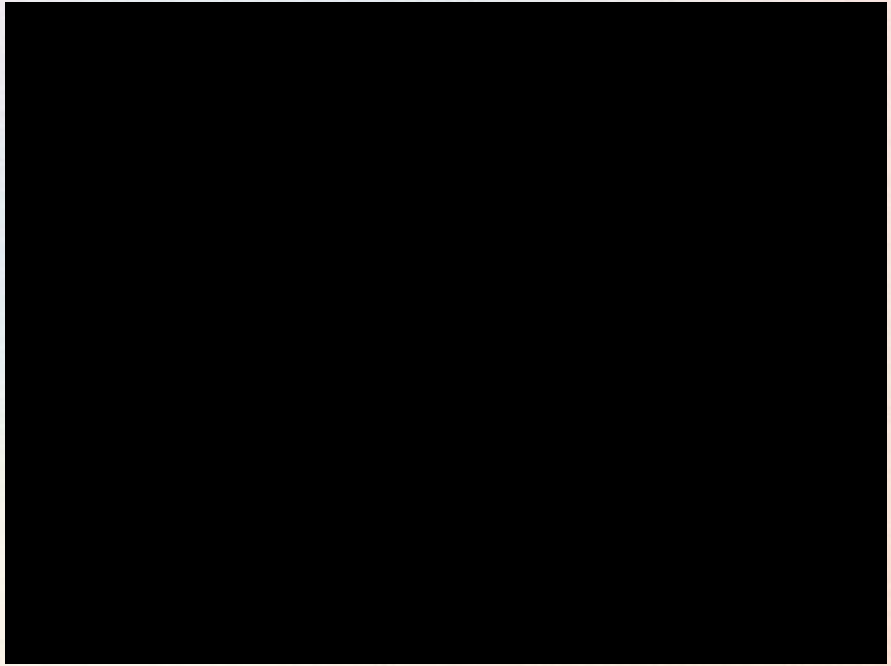
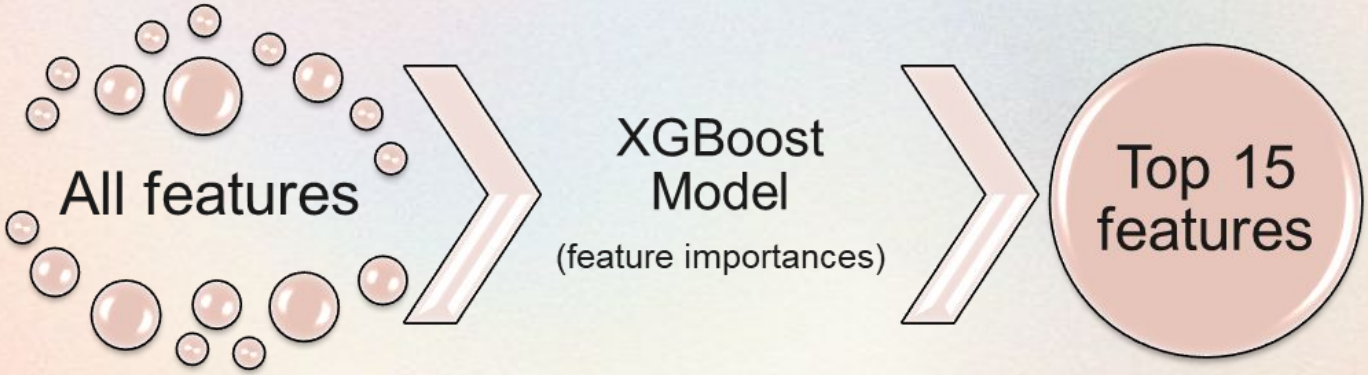
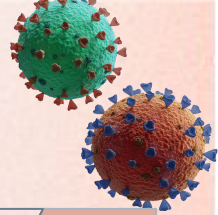


TABLE OF CONTENTS

Feature Selection



04 FEATURE IMPORTANCE: H1N1



Whether the respondent has health insurance

Does the respondent rent or own their home

Whether the respondent avoids large gatherings

How effective the respondent perceives the vaccine to be

Whether the doctor recommended the vaccine

Whether the respondent avoids touching their face

How risky the respondent perceives the vaccine to be

Respondent's employment industry

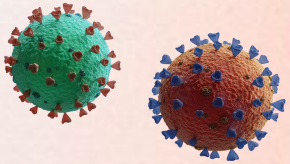
Whether the respondent avoids members of other households.

How much does the respondent know about the H1N1 flu

The marital status of the respondent

Whether the respondent avoids large gatherings.

04 FEATURE IMPORTANCE: Seasonal Flu



Whether the respondent has health insurance

Does the respondent rent or own their home

Whether the respondent has a chronic medical condition

How effective the respondent perceives the vaccine to be

Whether the doctor recommended the vaccine

Respondent's age

How risky the respondent perceives the vaccine to be

Respondent's employment industry **and status**

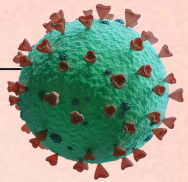
Whether the respondent is a health worker or works in a hospital

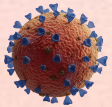
How much does the respondent know about the H1N1 flu

The highest level of education achieved by respondent

Whether the respondent fears getting sick from the vaccine

TABLE OF CONTENTS





05 PRODUCTION MODEL: H1N1

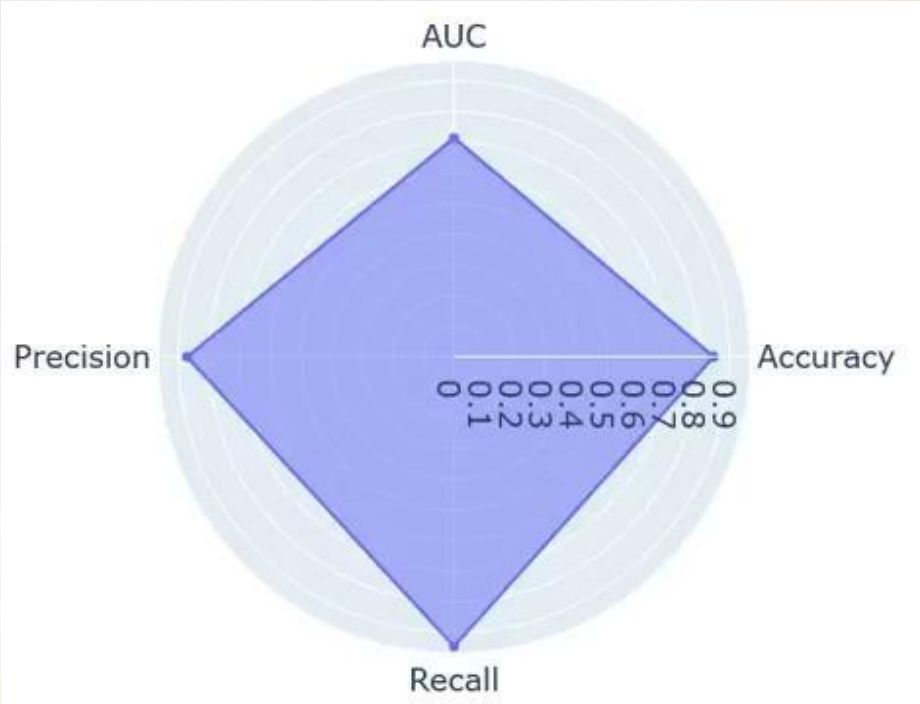
We trained a new XGBoost model just on the 15 selected features.

XGBoost Model on all features

Accuracy: 0.8518
AUC: 0.730
Precision: 0.879
Recall: 0.942

XGBoost Model on 15 selected features

Accuracy: 0.847
AUC: 0.714
Precision: 0.871
Recall: 0.945



05 PRODUCTION MODEL: Seasonal Vaccine

We trained a new XGBoost model just on the 15 selected features.

XGBoost Model on all features

Accuracy: 0.797
AUC: 0.794
Precision: 0.800
Recall: 0.825

XGBoost Model on 15 selected features

Accuracy: 0.849
AUC: 0.717
Precision: 0.872
Recall: 0.946

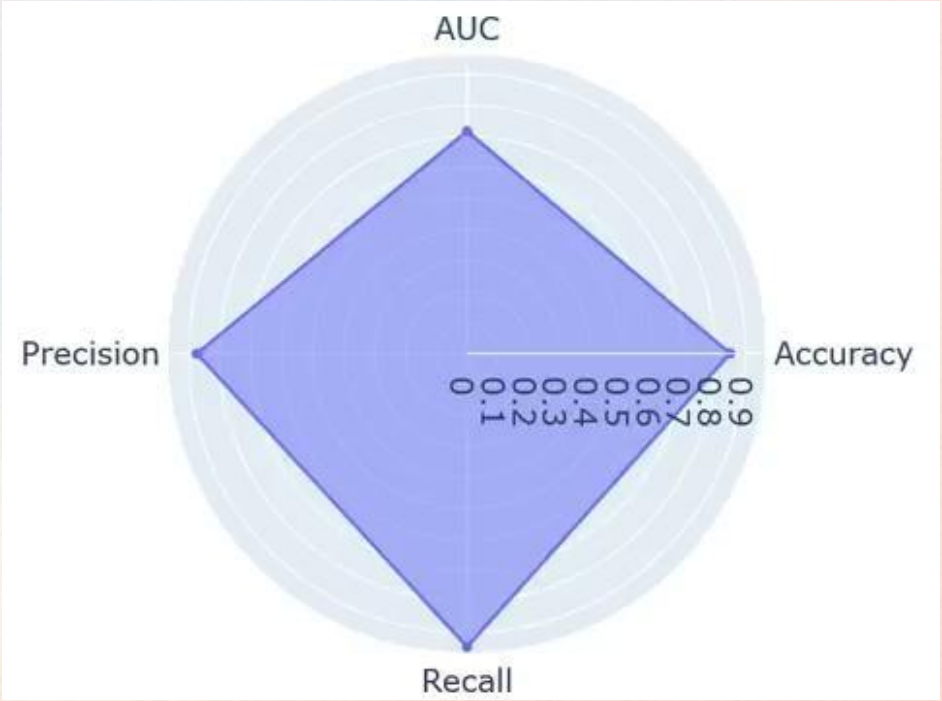
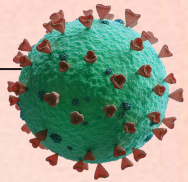
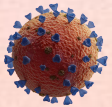


TABLE OF CONTENTS



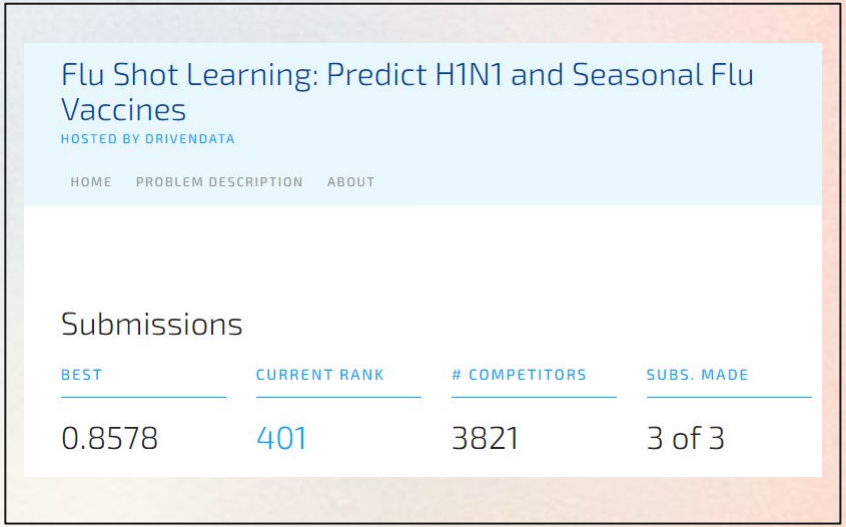


06 ADDITIONAL CONSIDERATIONS

The data set was obtained from an ongoing Data Science competition on *DrivenData.com*.

The metric used for the competition is AUC.

We were able to submit estimates from our Neural Network with AUC metric of 0.8578 and the current leading AUC metric is 0.8658.

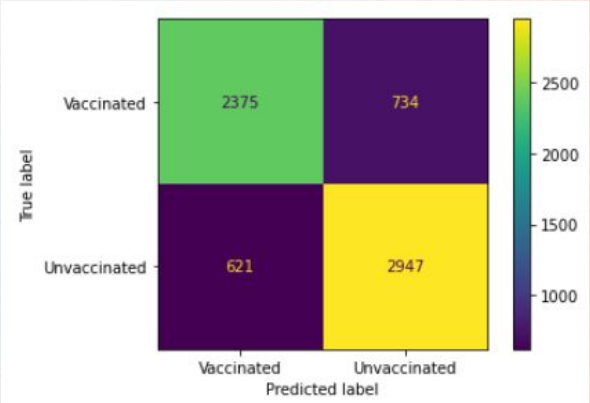


06 CONCLUSIONS

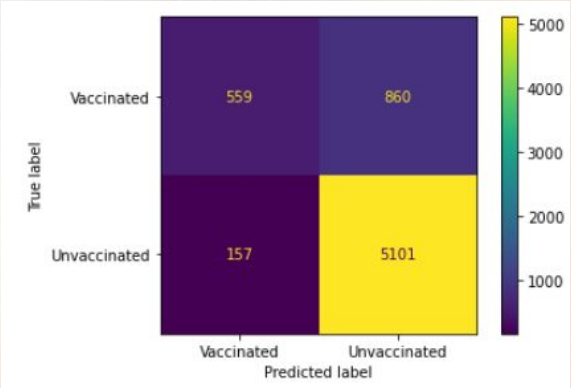
A better understanding of how these characteristics are associated with personal vaccination patterns can provide guidance for future public health efforts.

Goal 1: Predict which respondents refused vaccination.
We compared multiple optimized models and were able to make predictictions with a recall of

- 0.94 for H1N1, and
- 0.83 for seasonal flu



Seasonal Flu Vaccine

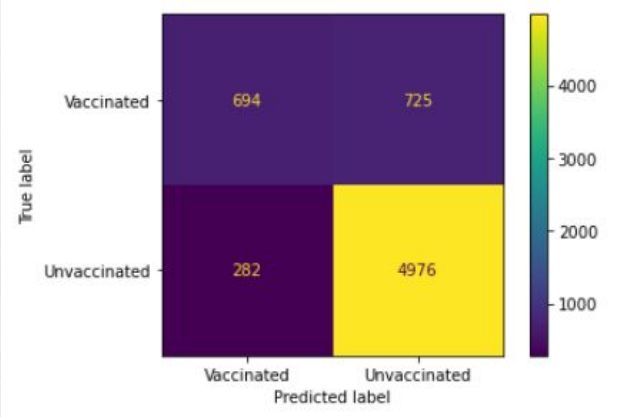


H1N1 Vaccine

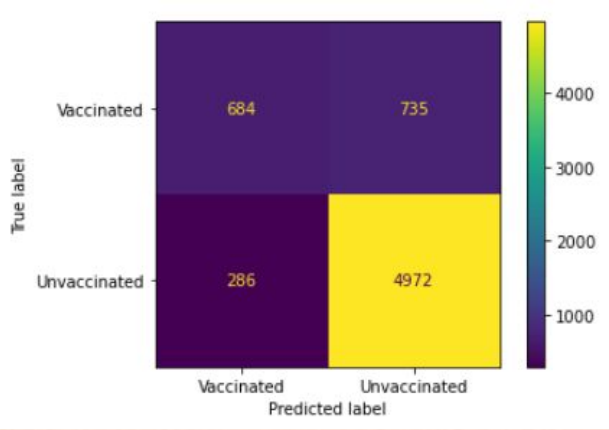
06 CONCLUSIONS

Goal 2: identify the features in the dataset that were most predictive of whether a person refused vaccination.

Using 15 of 35 features, our models predict whether or a person was unvaccinated with a recall of 0.94 for both H1N1 and seasonal flu vaccines.

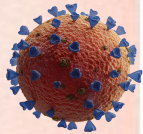


Seasonal Flu Vaccine



H1N1 Vaccine

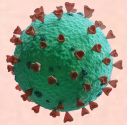
06 RECOMMENDATIONS



Goal 3: Recommend areas of focus to shorten the survey

Our models identified the most predictive features as:

- availability of health insurance,
- perception of the vaccine (effectiveness, risks),
- knowledge of H1N1,
- kind of dwelling,
- doctor's recommendation,
- employment,
- marital status,
- Behaviors,
- level of education,
- overall health
- age.



THANKS!

Do you have any questions?