H1N1 Flu Vaccination Status Prediction

Haotian Yang, Yuelin Ou

Our Data:

Source: National 2009 H1N1 Flu Survey (NHFS)

Timeline: 2009 Feb: Pandemic Started
2009 Oct: Vaccine Available
Early 2010: Survey conducted

Sample Size: 26707 respondents

Features: 35 (Numerical: 23, Categorical: 12), including personal background and hygiene.

Labels: 2 (H1N1 vaccination status, Seasonal Flu vaccination status)

Our GOAL:

1. Predict the likelihood of individuals receiving their H1N1 and seasonal flu vaccines.

2. Understand how personal factors impact the vaccination status and visualize the patterns

Overview

1. Data Preprocessing and Exploration

2. Model Training and Tuning

3. Conclusion

Step 1: Data Preprocessing and Exploration

Vaccination Status Proportion

H1N1:

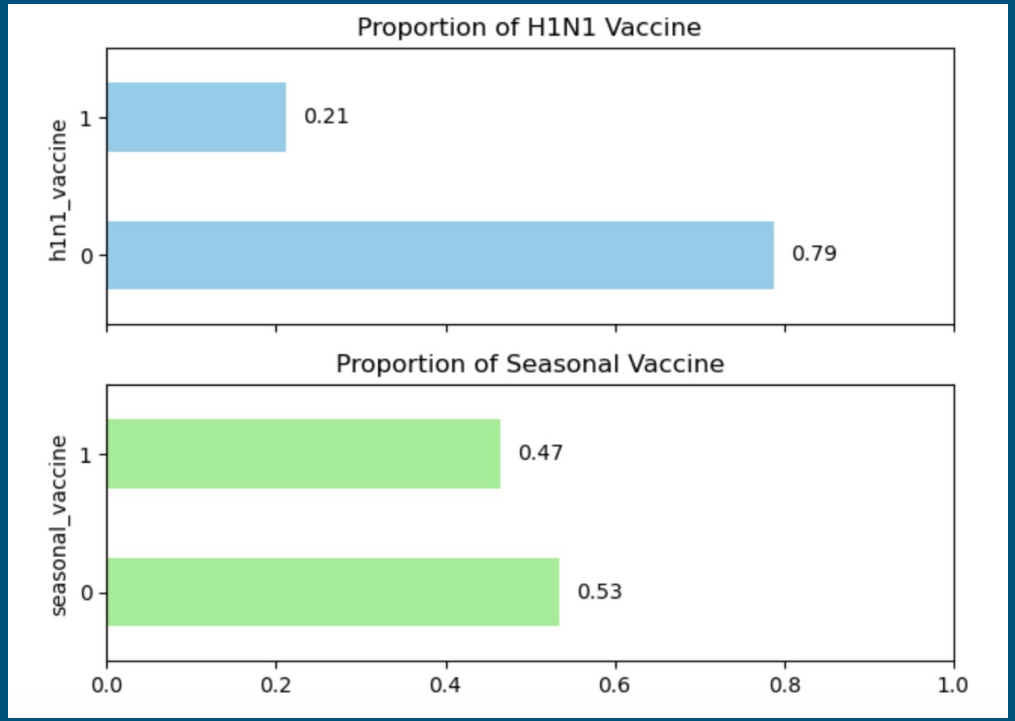
Taken: 21%

Not Taken: 79%

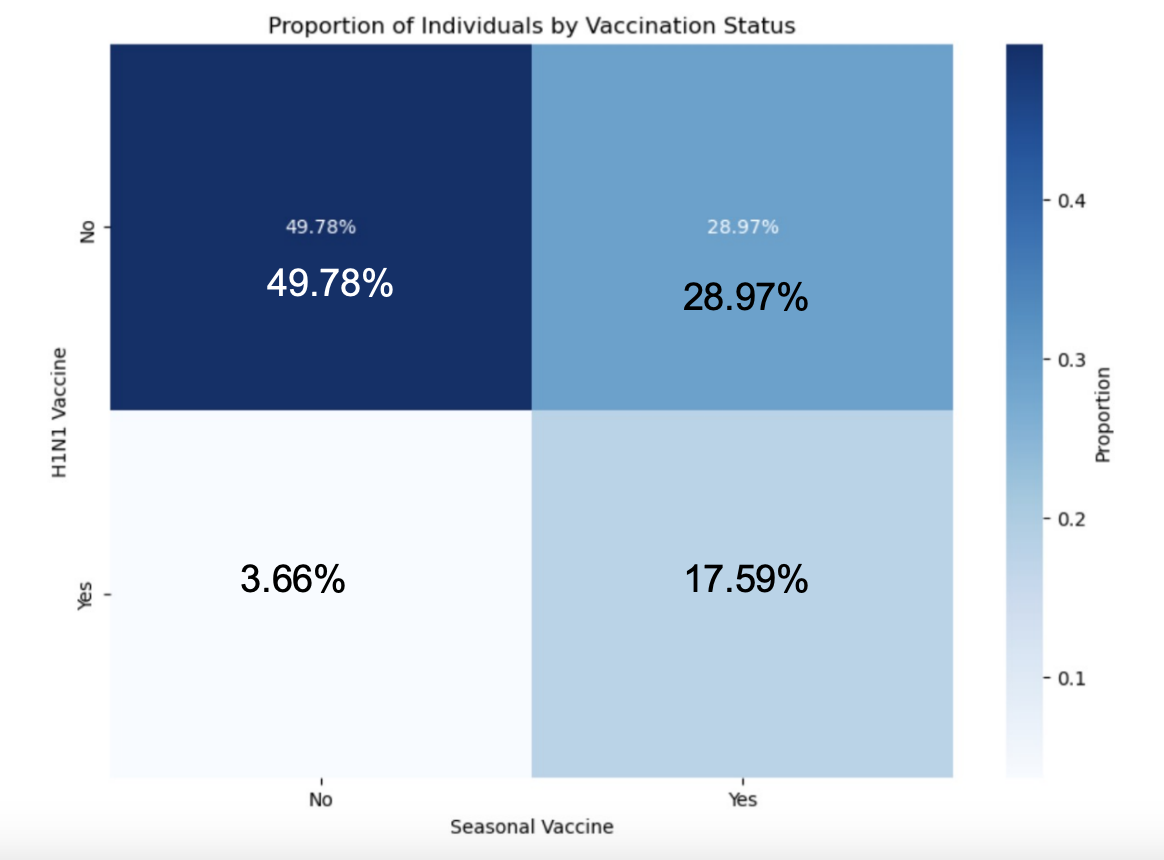
Seasonal Flu:

Taken: 47%

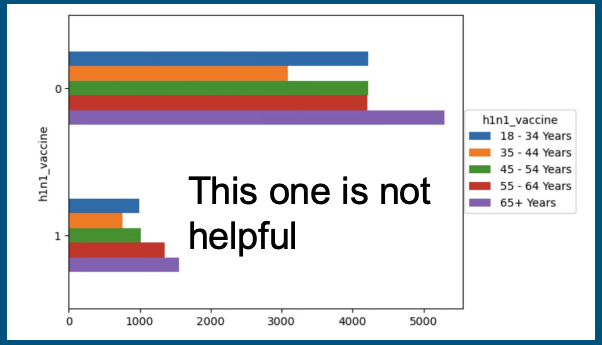
Not Taken: 53%



Vaccination Status Matrix

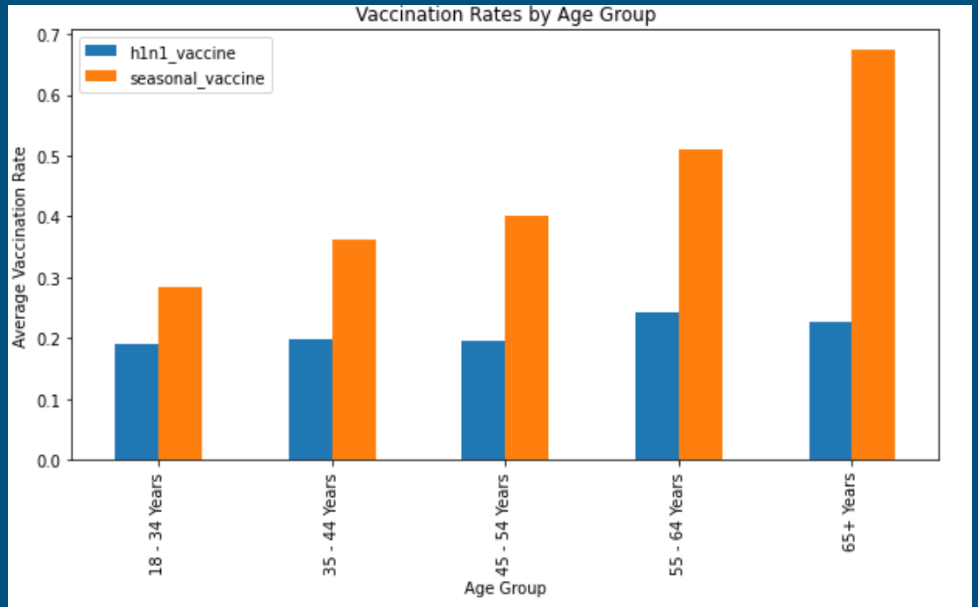


Influence of Each Feature



We analyzed data to observe how vaccination status varied across different features.

For instance, in the category of "age_group," we found that older individuals **are more likely** to get vaccinated against the seasonal flu. However, age **does not** have a clear influence on H1N1 vaccination rates.



Apply to All Features

Major features impacting H1N1 vaccination include concern level, knowledge, facemask behavior, handwashing behavior, doctor recommendations, health insurance, opinions, occupation, and industry.

Major features influencing seasonal flu vaccination are similar to those for H1N1 vaccination. However, **there are differences**, such as the effect of age.

Handling Missing Values

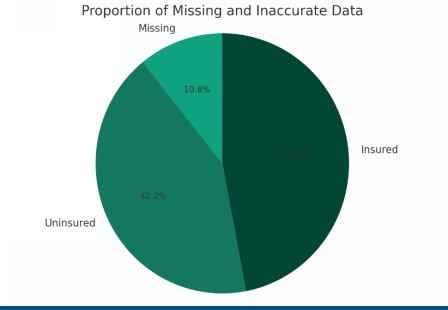
Over half of the feature columns had **missing values**, so we employed various methods to impute these values.

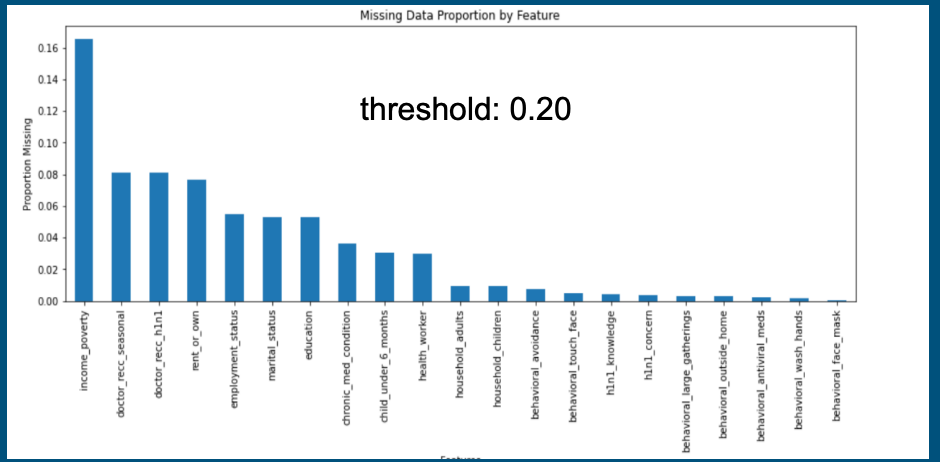
Initially, we set a threshold of 0.2 for retaining features, which means we considered dropping any column with more than 20% missing values. Three columns met this criterion: health insurance, employment industry, and employment occupation.

For imputation, we used strategies such as replacing missing values with the **mean, median, or mode**. Additionally, we designated **'unknown'** as a special category for unrecognized data.

```
h1n1_concern: 92 missing values
h1n1_knowledge: 116 missing values
behavioral_antiviral_meds: 71 missing values
behavioral_avoidance: 208 missing values
behavioral_face_mask: 19 missing values
behavioral_wash_hands: 42 missing values
behavioral_large_gatherings: 87 missing values
behavioral_outside_home: 82 missing values
behavioral_touch_face: 128 missing values
doctor_recc_h1n1: 2160 missing values
doctor_recc_seasonal: 2160 missing values
chronic_med_condition: 971 missing values
child_under_6_months: 820 missing values
health_worker: 804 missing values
health_insurance: 12274 missing values
opinion_h1n1_vacc_effective: 391 missing values
opinion_h1n1_risk: 388 missing values
opinion_h1n1_sick_from_vacc: 395 missing values
opinion_seas_vacc_effective: 462 missing values
opinion_seas_risk: 514 missing values
opinion_seas_sick_from_vacc: 537 missing values
age_group: 0 missing values
education: 1407 missing values
race: 0 missing values
sex: 0 missing values
income_poverty: 4423 missing values
marital_status: 1408 missing values
rent_or_own: 2042 missing values
employment_status: 1463 missing values
hhs_geo_region: 0 missing values
census_msa: 0 missing values
household_adults: 249 missing values
household_children: 249 missing values
employment_industry: 13330 missing values
employment_occupation: 13470 missing values
```

missing values statistics





Processing Categorical Data

As we observed, some features have two categories, allowing us to treat them as binary features.

To handle this, we introduced **One-Hot Encoding** to represent these binary features. Missing values within these categories were imputed using the **mode**.

For example, for the gender feature, we created binary columns such as 'is_Female' and 'is_Male', among others.

Now I have some general questions. I am required to ask this. Just to confirm, are you male or female?

(1) MALE **[SKIP TO Q87]**
(2) FEMALE
(99) REFUSED

Sample Encoding

Convert text labels into a machine–readable form.

Techniques such as One–Hot Encoding and Label Encoding are explained.

Example of encoding 'employment_status': Employed, Unemployed, Self–Employed.

Example:

Employment Status
Employed
Self-Employed
Unemployed

Encoded Value
0
1
2

Processing Categorical Data

However, the major problem with One-Hot Encoding is that it significantly **increases dimensionality**.

Therefore, for ordinal variables, we opt for **Ordinal Encoding**, and missing values are imputed using the **median**.

For example, for the variable 'Education', the encoding would be as follows:
1 for less than 12 years; 2 for 12 years; 3 for College; 4 for Graduate.

Processing Categorical Data

We also use Frequency Encoding, and address missing values by adding a category called "unknown" before encoding.

$$FrequencyEncoding = \frac{frequency(category)}{size(data)}$$

For example:

Race

Processing Numerical Data

For numerical features, we don't need extensive convertion.

There are 4 major types of numerical features.

1. Opinion: 1– 5 (Missing values: median)
2. Concern: 1–5 (Missing values: median)
3. Yes or no: (Missing values: mode)
4. How many: (Missing values: median)

Now I'm going to ask you your opinion about the H1N1 flu vaccine. When the H1N1 flu vaccine is available this fall, how likely would you be to get this vaccination? Would you say that you are very likely, somewhat likely, somewhat unlikely, or very unlikely to get this vaccination?

(1) VERY LIKELY	**[SKIP TO Q23]**
(2) SOMEWHAT LIKELY	**[SKIP TO Q23]**
(3) SOMEWHAT UNLIKELY	**[SKIP TO Q23]**
(4) VERY UNLIKELY	**[SKIP TO Q23]**
(77) DON'T KNOW	**[SKIP TO Q23]**
(99) REFUSED	**[SKIP TO Q23]**

How concerned are you about the H1N1 flu? Would you say you are very concerned, somewhat concerned, not very concerned, or not at all concerned?

(1) VERY CONCERNED
(2) SOMEWHAT CONCERNED
(3) NOT VERY CONCERNED
(4) NOT AT ALL CONCERNED
(77) DON'T KNOW
(99) REFUSED

Frequent hand washing or use of hand sanitizer.
READ IF NECESSARY: Have you done this as a result of this current pandemic?

(1) YES
(2) NO
(77) DON'T KNOW
(99) REFUSED

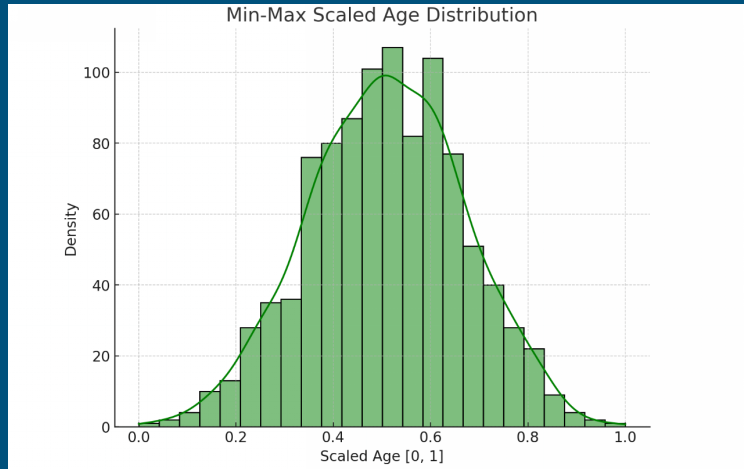
What is the actual number of people under 18 in your household?
_____ CHILDREN **[IF CHILDREN = 0 SKIP TO Q87]**

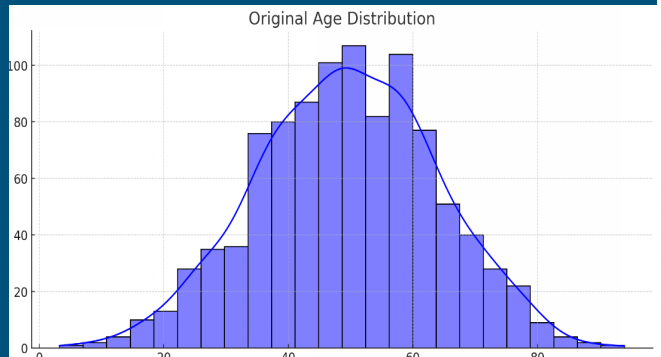
Normalizing Numerical Features

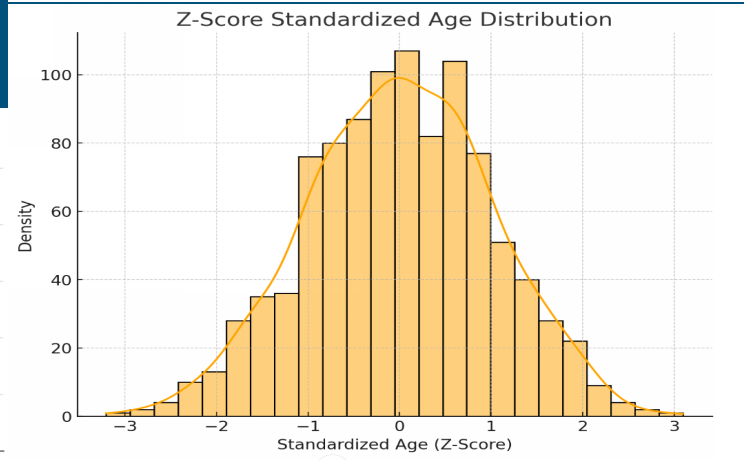
Why scale adjustment is crucial for model performance.

Methods: Min–Max Scaling and Z-Score Standardization.

Result: Balanced influence of each feature in the predictive model.



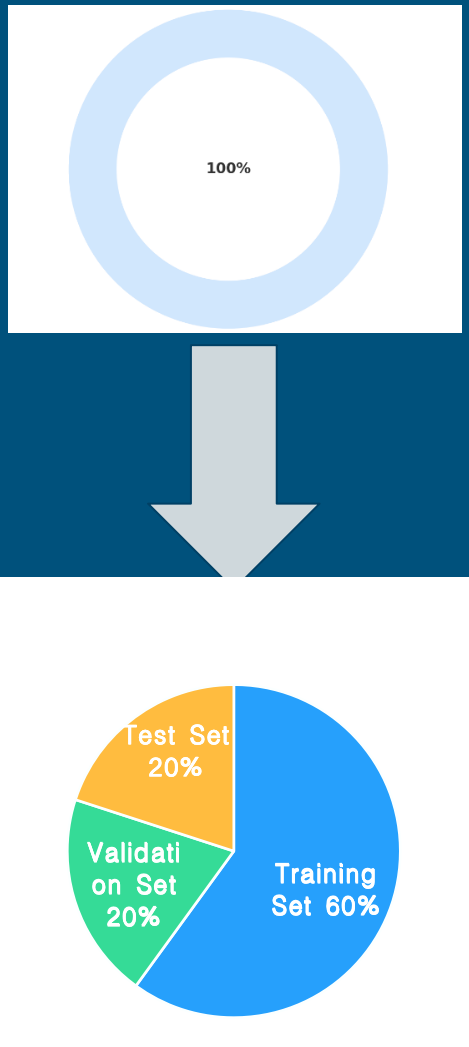




Step 2: Model Training and Tuning

Training Set Praperation:

We split the original dataset and created a training set, a validation set, and a test set in the ratio of **6:2:2** to avoid overfitting and test the model's accuracy.

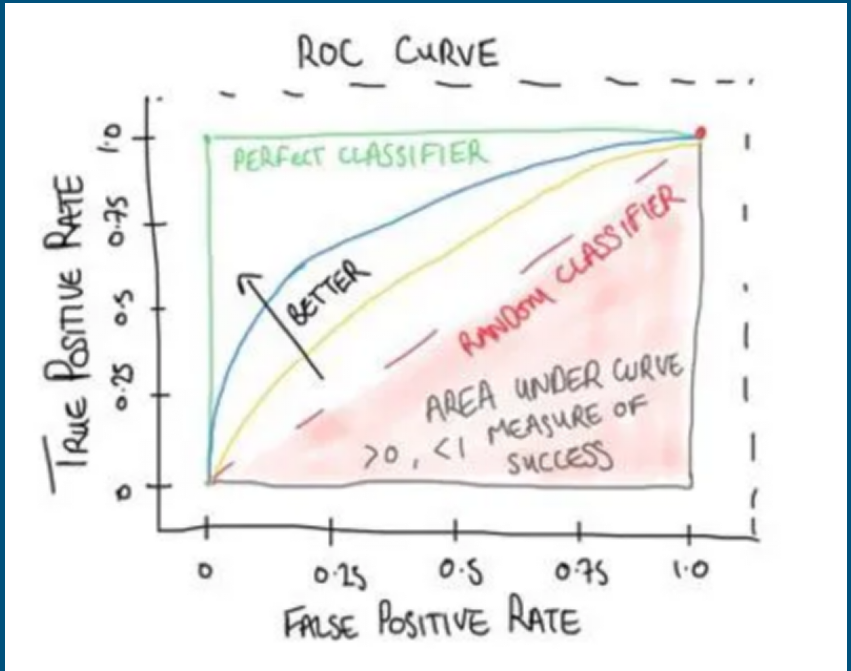


Model selection

1. Logistic Regression (Benchmark)
2. KNN (Euclidean Distance)
3. Decision Tree (basic)
4. GBDT (XGBoost Improvement)

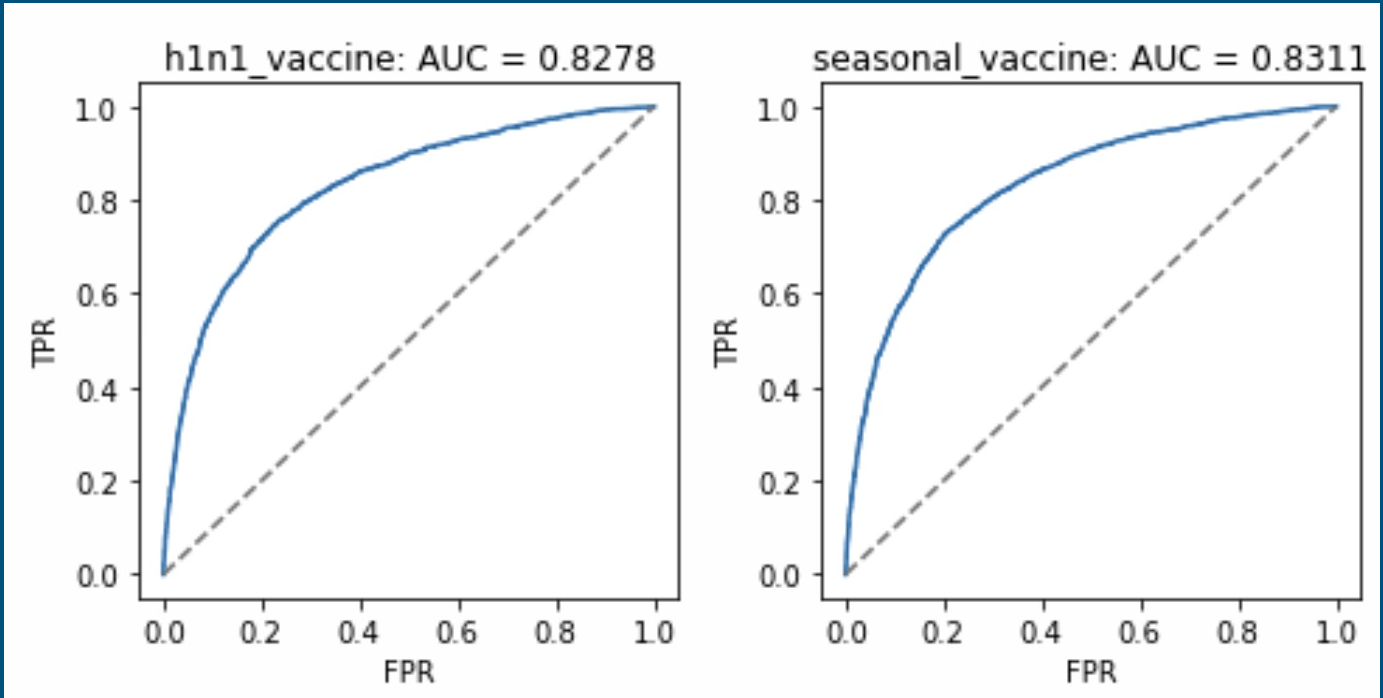
Evaluation Method

ROC curve (receiver operating characteristic curve)

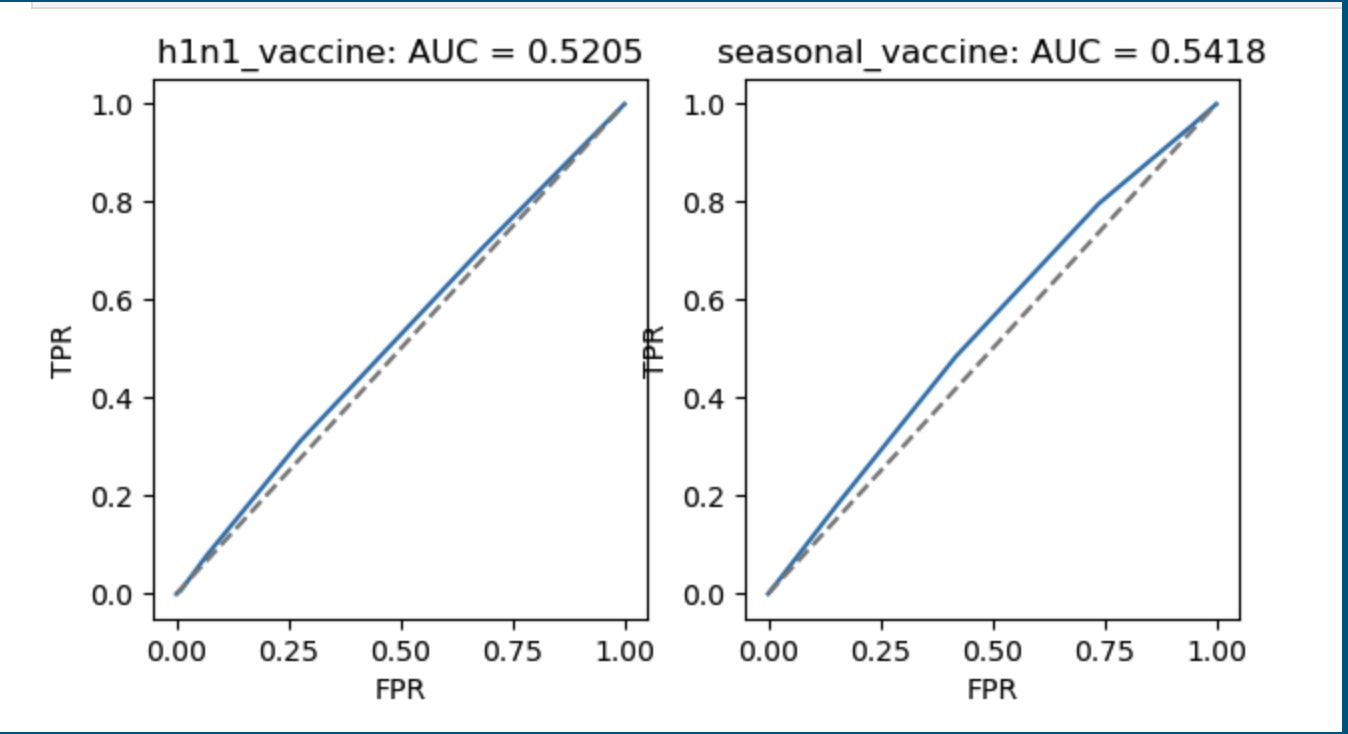


Becnmark

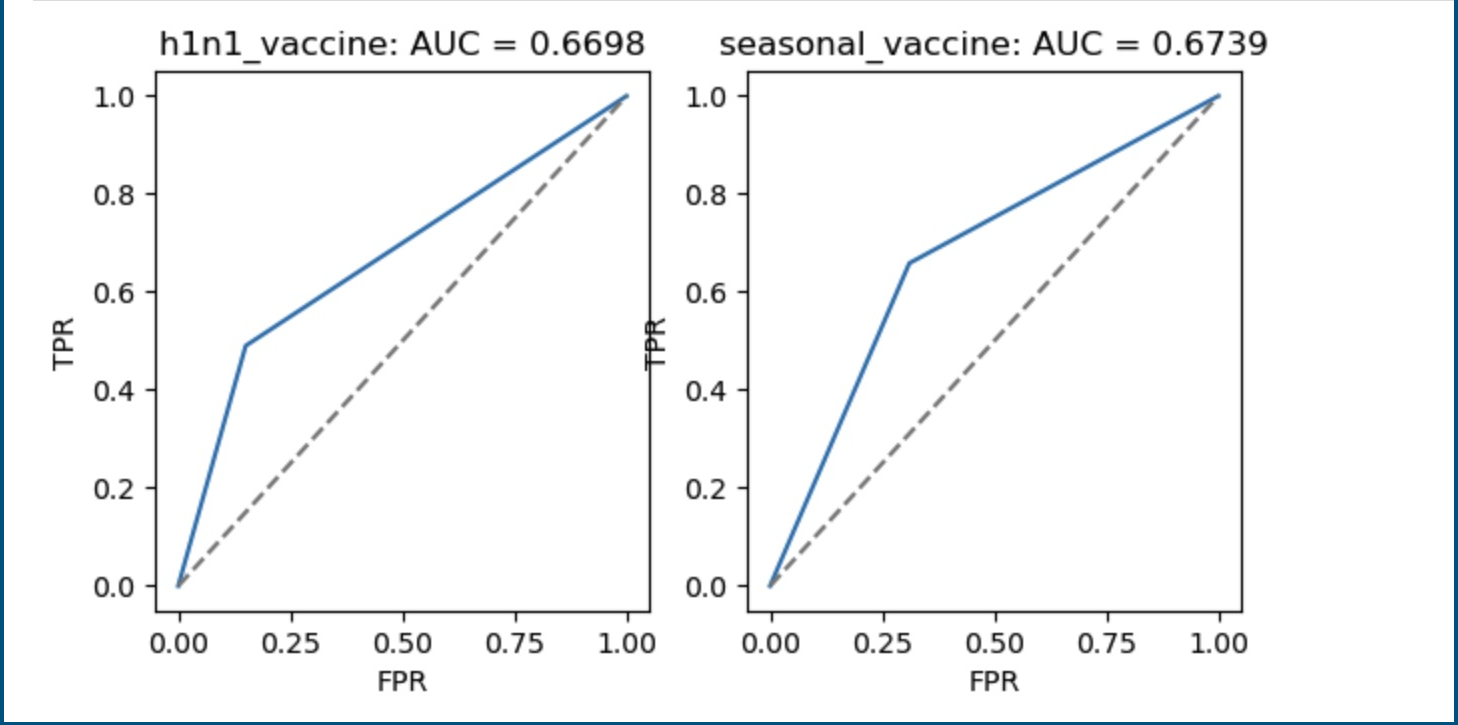
Logistic Regression



KNN (Euclidean Distance)



Decision Tree

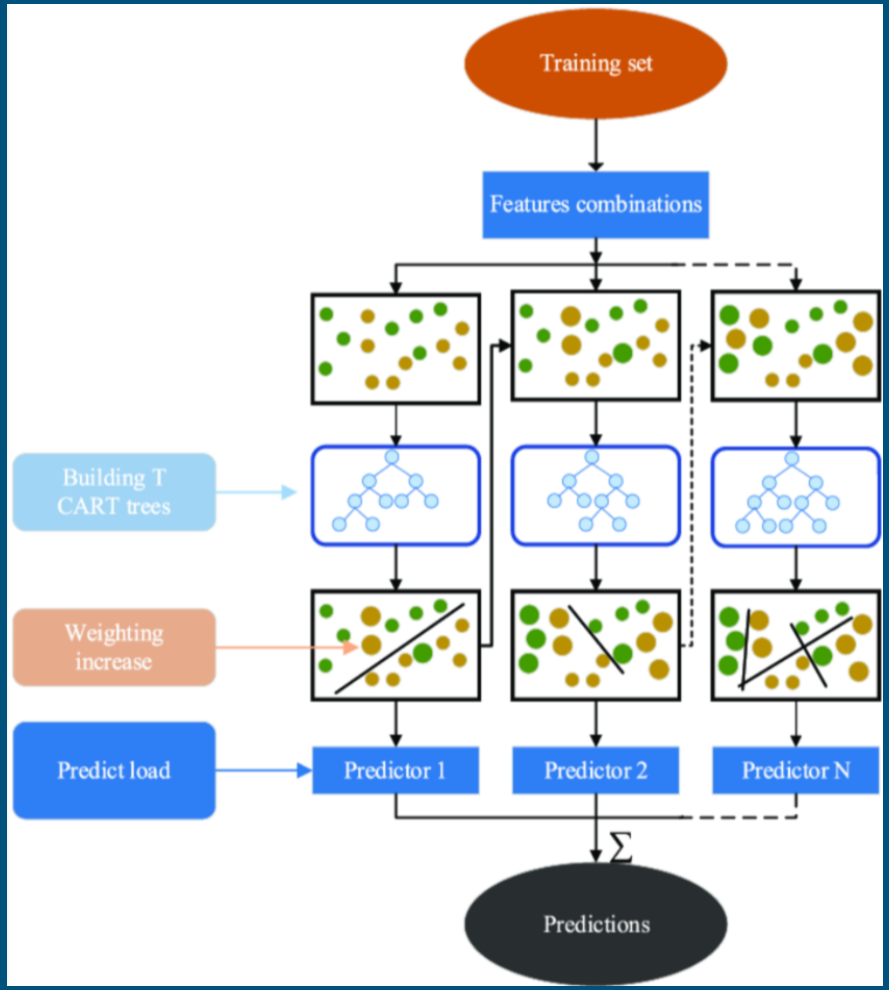


Final Model Selection

- XGBoost Classifier

Gradient Boosting Decision Tree, GBDT

Gradient Boosted Decision Trees (GBDT) is a machine learning technique that builds a powerful predictive model by combining multiple simple models (usually decision trees).



Tuning:

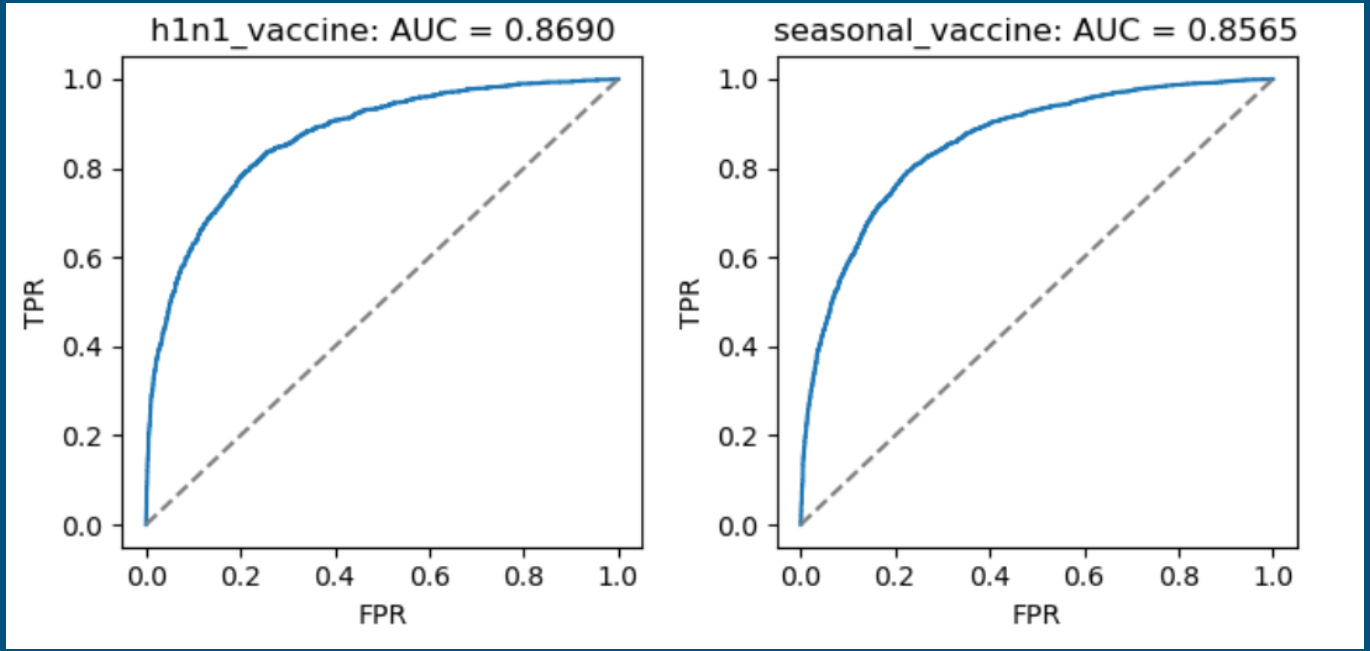
```
param_grid = {
    'estimator__n_estimators': [100, 200, 300],
    'estimator__learning_rate': [0.01, 0.05, 0.1],
    'estimator__max_depth': [3, 4, 5],
    'estimator__subsample': [0.7, 0.8, 0.9],
    'estimator__colsample_bytree': [0.7, 0.8, 0.9],
}
```

Hyperparameter tuning is crucial for enhancing model accuracy, mitigating overfitting, and maximizing predictive performance.

XGBoost is an ensemble learning algorithm that sequentially adds multiple decision trees to form a model, using their combined predictions to minimize prediction errors and enhance overall accuracy.

XGBoost introduces a fine-grained tree segmentation algorithm (i.e., approximation algorithm) that can handle larger data sets and provides tools for model tuning and interpretive analysis.

Final Evaluation



The AUC for h1n1 vaccine prediction is 0.8690, indicating high accuracy, while the AUC for seasonal vaccine prediction is slightly lower at 0.8565, yet still showing strong predictive ability. Both values suggest the models are well–calibrated and significantly better than random guessing.

Step 3: Conclusion

What Can We Improve?

1. Methods proccessing the missing values can be improved by more comparison.

2. We can optimize the data processing procedures by utilizing transformer which can better customize the input data.

3. Try to tune the models with more parameters.

4. Visualization can be integrated to make more straightforward comparison.

How this Model works?

1. The model could be reused by inputting other pandemics' data to compare the results and gain some insights on the vaccination realization changes if similar surveys can be conducted.

2. The model could help people understand how people think about vaccination in with different backgrounds.

Thank you for listening our presentation!