

VACCINE PREDICTION PROJECT

data report on the prediction of vaccine uptake by individuals

Linda Temoet | Data Scientist

# BUSINESS UNDERSTANDING

## OVERVIEW

H1N1, commonly referred to as swine flu, is a subtype of the influenza A virus that can infect humans, birds, and pigs. The virus gained widespread attention in 2009 when a new strain emerged, leading to the first influenza pandemic in over four decades. This novel H1N1 strain was particularly concerning due to its rapid human-to-human transmission and its ability to infect individuals across a wide age range, including those without preexisting conditions.

During the pandemic, the World Health Organization (WHO) declared H1N1 a global health emergency. Vaccination efforts were a critical component of the public health response, aimed at mitigating severe cases, hospitalizations, and fatalities. Despite these efforts, vaccine uptake varied significantly across populations due to a combination of logistical challenges, misinformation, and vaccine hesitancy.

Although the pandemic officially ended in August 2010, H1N1 has since become a seasonal flu virus, contributing to annual influenza epidemics. Understanding the factors influencing vaccine acceptance remains a priority for public health authorities to improve vaccination rates and enhance preparedness for future outbreaks

## PROBLEM STATEMENT

Vaccination rates are lower than desired, leading to preventable disease outbreaks.

## OBJECTIVES

* To analyze which demographic aspects (e.g.age, education) most influence vaccine uptake.
* To evaluate the attitudinal features affecting vaccine use.
* To build a machine learning model to predict whether an individual will take a vaccine based on demographic and attitudinal factors.

## SUCCESS CRITERIA

* Deliver a report highlighting the top 10 predictors.
* AUC at least 0.80 to show the difference between the two groups

# DATA UNDERSTANDING

## OVERVIEW

The data was obtained from [https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/](https://www.google.com/url?q=https%3A%2F%2Fwww.drivendata.org%2Fcompetitions%2F66%2Fflu-shot-learning%2Fpage%2F210%2F) Originally provided courtesy of the United States National Center for Health Statistics. U.S. Department of Health and Human Services (DHHS). National Center for Health Statistics. The National 2009 H1N1 Flu Survey. Hyattsville, MD: Centers for Disease Control and Prevention, 2012.

The data represents statistics on vaccine uptakes of h1n1 vaccine (swine flu) and the seasonal flu vaccine in various regions in the US.

Each record is for a respondent for the attitudinal and demographic features and whether the have had their seasonal and h1n1 vaccination

The data came in 2 CSV files, one for the independent variables(X) and the other for the target variable. The datasets were loaded and merged into one dataframe first, before work on them began.

The dataframe was then checked for the shape, info, column names and describe. This gave insight into the data;

The number of columns and rows (rows = 26,707 and columns were 37)

The column names were as follows

### Attitudinal Factors:

* **h1n1\_concern:** Individual's level of concern about the H1N1 virus.
* **h1n1\_knowledge:** Individual's level of knowledge about the H1N1 virus.
* **behavioral\_antiviral\_meds:** Individual's intention to use antiviral medications.
* **behavioral\_avoidance:** Individual's intention to avoid sick people.
* **behavioral\_face\_mask:** Individual's intention to wear a face mask.
* **behavioral\_wash\_hands:** Individual's intention to frequently wash hands.
* **behavioral\_large\_gatherings:** Individual's intention to avoid large gatherings.
* **behavioral\_outside\_home:** Individual's intention to limit time spent outside the home.
* **behavioral\_touch\_face:** Individual's intention to avoid touching their face.
* **doctor\_recc\_h1n1:** Whether the doctor recommended the H1N1 vaccine.
* **doctor\_recc\_seasonal:** Whether the doctor recommended the seasonal flu vaccine.
* **opinion\_h1n1\_vacc\_effective:** Individual's opinion on the effectiveness of the H1N1 vaccine.
* **opinion\_h1n1\_risk:** Individual's perception of their risk of getting sick from H1N1.
* **opinion\_h1n1\_sick\_from\_vacc:** Individual's belief that they will get sick from the H1N1 vaccine.
* **opinion\_seas\_vacc\_effective:** Individual's opinion on the effectiveness of the seasonal flu vaccine.
* **opinion\_seas\_risk:** Individual's perception of their risk of getting sick from seasonal flu.
* **opinion\_seas\_sick\_from\_vacc:** Individual's belief that they will get sick from the seasonal flu vaccine.

### Demographic Factors:

* **age\_group:** Individual's age group.
* **education:** Individual's education level.
* **race:** Individual's race.
* **sex:** Individual's sex.
* **income\_poverty:** Individual's income level relative to the poverty level.
* **marital\_status:** Individual's marital status.
* **rent\_or\_own:** Whether the individual rents or owns their home.
* **employment\_status:** Individual's employment status.
* **hhs\_geo\_region:** Geographic region of the individual's residence.
* **census\_msa:** Metropolitan Statistical Area of the individual's residence.
* **household\_adults:** Number of adults in the household.
* **household\_children:** Number of children in the household.
* **employment\_industry:** Individual's industry of employment.
* **employment\_occupation:** Individual's occupation.

### Target Variables:

* **h1n1\_vaccine:** Whether the individual received the H1N1 vaccine.
* **seasonal\_vaccine:** Whether the individual received the seasonal flu vaccine.

The datatypes were as follows:

0 h1n1\_concern 26615 non-null float64

1 h1n1\_knowledge 26591 non-null float64

2 behavioral\_antiviral\_meds 26636 non-null float64

3 behavioral\_avoidance 26499 non-null float64

4 behavioral\_face\_mask 26688 non-null float64

5 behavioral\_wash\_hands 26665 non-null float64

6 behavioral\_large\_gatherings 26620 non-null float64

7 behavioral\_outside\_home 26625 non-null float64

8 behavioral\_touch\_face 26579 non-null float64

9 doctor\_recc\_h1n1 24547 non-null float64

10 doctor\_recc\_seasonal 24547 non-null float64

11 chronic\_med\_condition 25736 non-null float64

12 child\_under\_6\_months 25887 non-null float64

13 health\_worker 25903 non-null float64

14 health\_insurance 14433 non-null float64

15 opinion\_h1n1\_vacc\_effective 26316 non-null float64

16 opinion\_h1n1\_risk 26319 non-null float64

17 opinion\_h1n1\_sick\_from\_vacc 26312 non-null float64

18 opinion\_seas\_vacc\_effective 26245 non-null float64

19 opinion\_seas\_risk 26193 non-null float64

20 opinion\_seas\_sick\_from\_vacc 26170 non-null float64

21 age\_group 26707 non-null object

22 education 25300 non-null object

23 race 26707 non-null object

24 sex 26707 non-null object

25 income\_poverty 22284 non-null object

26 marital\_status 25299 non-null object

27 rent\_or\_own 24665 non-null object

28 employment\_status 25244 non-null object

29 hhs\_geo\_region 26707 non-null object

30 census\_msa 26707 non-null object

31 household\_adults 26458 non-null float64

32 household\_children 26458 non-null float64

33 employment\_industry 13377 non-null object

34 employment\_occupation 13237 non-null object

35 h1n1\_vaccine 26707 non-null int64

36 seasonal\_vaccine 26707 non-null int64

## DATA PREPARATION

The data was checked for missing and duplicated values. No duplicates were found but a lot of missing values were found.

The columns containing missing values were then checked for the values contained therein.

The columns hhs\_geo\_region,census\_msa,employment\_industry, employment\_occupation were then dropped because they contained values that were not easily understood.

The health insurance column was important so the missing values were imputed with the median value.

The rest of the rows with missing values were dropped. This reduced the records to 19642 which is still considered a good dataframe number for analysis

## EXPLORATORY DATA ANALYSIS

**Analyzing Demographics:**

A list was defined named demographic containing features related to individual characteristics like race, sex, education, etc. The target variable was set as h1n1\_vaccine, indicating whether the individual received the H1N1 vaccine.

A count plot was made which visualized the distribution of each demographic category (y-axis) according to whether the individual received the H1N1 vaccine (hue).

**Analyzing Attitudinal Factors:**

A list was defined named attitudinal containing features related to attitudes and beliefs about H1N1 and vaccination with h1n1\_vaccine being the target

pd.crosstab was used to create a contingency table. This table showed the cross-tabulation between the attitudinal factor and the target variable. It counted how many individuals fall into each category based on their attitude and vaccination status.

A stacked bar chart was made which then visualized the distribution of the attitudinal factor categories on the x-axis, with separate bars for vaccinated and unvaccinated individuals (stacked).

These were the findings:

Demographic Factors and Vaccine Uptake

Based on the visualizations, we can observe several trends:

* Age: Older individuals tend to have higher vaccination rates. This could be due to increased risk perception, stronger healthcare seeking behaviors, or historical experiences with influenza.
* Income: Higher income individuals are more likely to be vaccinated. This could be due to better access to healthcare, health insurance, and health information.
* Education: Higher levels of education are associated with higher vaccination rates. This could be due to better health literacy and understanding of the benefits of vaccination.
* Marital Status: Married individuals tend to have lower vaccination rates. This could be due to family responsibilities or other factors that might influence their decision-making.
* Race and Ethnicity: Racial and ethnic minorities may have lower vaccination rates due to historical mistrust of the healthcare system, lack of access to healthcare, or cultural beliefs.

Attitudinal Factors and Vaccine Uptake

* Perceived Risk: Individuals who perceive a higher risk of H1N1 infection are more likely to get vaccinated.
* Knowledge: Individuals with higher levels of knowledge about H1N1 are more likely to get vaccinated.
* Behavioral Intentions: Individuals who intend to engage in preventive behaviors, such as using antiviral medications, wearing masks, or washing hands, are more likely to get vaccinated.
* Opinions: Perceived vaccine effectiveness and perceived risk of infection are positively associated with H1N1 vaccine uptake.
* Doctors recommendation: Individuals are more likely to take the vaccine if reccomended by a doctor

# MODELING

The model of choice was Logistic regression since the target variable was binary

Started by checking the distribution of the target variable. It was imbalanced so the decision was made to use SMOTE for oversampling the minority class

Then a baseline model was created of the attitudinal factors and how they affect vaccine uptake and the metrics were checked.

These were the metrics obtained:

* Accuracy (0.76): The model correctly predicts 76.15% of all cases
* Precision (0.48): Of all the cases the model predicted as vaccinated, only 48.4% were correct.
* Recall (0.70): The model correctly identifies 70.4% of those who are actually vaccinated.
* F1 Score (0.57): This metric balances precision and recall, with a score of 57.3%.
* **AUC Score (0.81):** The model's ability to distinguish between classes (vaccinated vs. not vaccinated) is 0.81%, which is reasonably good.

A new model 2 was made which contained the numeric columns only. The metrics improved significantly.

* Accuracy: the model is correct 80.04% of the time.
* Precision: when the model predicts a positive class, it's correct 54.25% of the time.
* Recall (Sensitivity): the model correctly identifies 78.52% of the actual positive cases.
* F1-Score: the F1-score is 0.6416, indicating a moderate balance between precision and recall.
* **AUC Score**: the AUC score is 0.864, which is a really good score to distinguish between the two groups.

A model 3 was then created which contained all the values in the dataframe (having the categorical columns dummy encoded). These were the metrics:

* Accuracy: the model is correct 79.84% of the time.
* Precision: This measures the proportion of positive predictions that are actually positive. In this case, when the model predicts a positive class, it's correct 53.96% of the time.
* Recall (Sensitivity): the model correctly identifies 77.55% of the actual positive cases.
* F1-Score: the F1-score is 0.6364, indicating a moderate balance between precision and recall.
* **AUC Score**: the AUC score is 0.8647, which is a decent score.

The model with the best metrics was model 2.

The decision tree classifier was attempted, but the metrics of choice were not very good.

Hyperparameter testing and feature selection were done; but model 2 with default parameters was just as good.

The features in model 2 were then checked for better understanding of the features affecting swine flu uptake.

Remembering that the metrics of choice was AUC, our target was achieved.

# CONCLUSION

From the given features, we can identify several key factors that significantly impact vaccine uptake:

* Attitudinal Factors: Individuals with higher perceived risk of H1N1 infection are more likely to get vaccinated.
* Belief in the effectiveness of the vaccine is a strong predictor of vaccination behavior.
* Concerns about side effects can negatively impact vaccine uptake.
* Behavioral Intentions: Individuals who intend to engage in preventive behaviors (e.g., wearing masks, washing hands) are more likely to get vaccinated.
* Social and Demographic Factors: Household size and the presence of children may influence vaccination decisions.
* Healthcare Access: Access to healthcare providers and insurance coverage can impact vaccine uptake.
* Improving the Model: While the model performs well, there are opportunities for further improvement.

While the current AUC is good, there are always opportunities for improvement:

1. Feature Engineering: Create new features by combining or transforming existing ones. Consider incorporating external data sources (e.g., socio-economic indicators, climate data) if relevant.
2. Model Selection: Consider alternative models like Random Forest, XGBoost, or LightGBM, which often perform well on complex datasets. Use ensemble methods to combine multiple models and improve overall performance.
3. Data Quality and Quantity: Gather more data to improve the model's generalization ability.

By continuously evaluating and refining the model, we can strive to improve its performance and gain deeper insights into the factors influencing H1N1 vaccine uptake.

# RECOMMENDATIONS

Based on the analysis of the H1N1 vaccine uptake model, the following recommendations can be made to the public health ministry:

1. Targeted Public Health Campaigns
2. Improve Access to Vaccination
3. Address Vaccine Hesitancy
4. Utilize Data-Driven Insights By implementing these recommendations, the public health ministry can effectively increase H1N1 vaccine uptake, protect public health, and mitigate the impact of future outbreaks.