**Predicting Vaccination Rates for H1N1 and Seasonal Flu:**

Exploring Factors that Encourage Vaccination Acceptance

Isabella Oakes

Applied Data Science Master’s Program

Shiley Marcos School of Engineering / University of San Diego  
 ioakes@sandiego.edu

Sarah Alqaysi

Applied Data Science Master’s Program  
 Shiley Marcos School of Engineering / University of San Diego  
 salqaysi@sandiego.edu

Omar Elfeky

Applied Data Science Master’s Program  
 Shiley Marcos School of Engineering / University of San Diego  
 oelfeky@sandiego.edu

**ABSTRACT**

It is commonly agreed that preventing a disease is much easier than treating it, whether viewed from a cost or health standpoint. Vaccination is one of the medical industry’s greatest innovations and by far the most effective way of preventing infectious diseases like the seasonal flu and H1N1 influenza. For most people, contracting a flu introduces mild and treatable symptoms; however, in many cases hospitalization becomes necessary leading to serious illness with many death cases for patients with chronical diseases as well as elderly people. While vaccination of H1N1 and seasonal flu are still considered optional, the focus of this study is to find out the likelihood of someone choosing to receive those optional vaccines.

**KEYWORDS**

*h1n1, influenza, vaccination, infectious diseases, health, pandemic, flu*

1**Introduction**

To illustrate the importance of flu vaccination, the Centers for Disease Control and Prevention (CDC; 2021b) published in 2019-2020 that the number of admitted and hospitalized patients was reduced by nearly 105,000 in flu-related cases. The CDC also cited a study where it was shown that people who chose to get the influenza vaccination “had a 31% reduced risk of death compared with unvaccinated patients” (Ferdinands et al., 2021, p. 3678). In addition, another study was cited by the CDC that claims “influenza vaccination was associated with a 59% reduction in the odds of ICU admission” (Thompson et al., 2018, p. 5916). They continue on to say that those patients also spent less time in the hospital, with an average of four days less. Despite new surges of the COVID-19 pandemic, vaccination has certainly aided in bringing life closer to normality. The focus of this paper will be on predicting how likely a person is to choose to receive a seasonal flu and H1N1 vaccine based on their responses to 35 questions. A model was created based on the data that was collected to predict the likelihood that a participant will receive the seasonal flu vaccine as well as the H1N1 vaccine. The model will give insights into the general likelihood of people receiving each vaccine as well as what impacts the decision the most. These insights can help inform future methods for educating and encouraging people to get vaccinated, which is vitally important as was seen in the COVID-19 pandemic.

2**Background**

The influenza virus has four broad categories: A, B, C, and D. The seasonal flu vaccine covers the anticipated most popular strains. Today, the quadrivalent seasonal flu contains two strains from each the A and B category (Center for Disease Control and Prevention, 2021a). One of the most prominent pandemics before COVID-19 was the 2009 H1N1 swine flu (Tosh, 2021). The United States and Europe have similar flu vaccination rates of 38% and 40% respectively, but there is significant variation regionally (Dinerstein, 2018). One notable difference between the seasonal flu and the 2009 swine flu pandemic was the age group of the deaths. The typical flu seasonal flu affects young children, those 65 and older, and pregnant women the most whereas pandemic influenza viruses in the past put young healthy people at risk (Centers for Disease Control and Prevention, 2021). The costs of getting the seasonal flu are direct, such as hospital visits, or indirect costs, such as a loss of productivity at work or school. Vaccines are an effective and cost-effective tool against the flu.

2.1**Lack of Vaccine Equity**

Vaccine hesitancy is a major public health concern. Comparing and contrasting historical vaccine uptake data for routine vaccines such as the seasonal flu and a pandemic vaccine uptake data such as swine flu could provide immense value especially in the event of more lethal and widespread pandemics occurring. Traditionally, minorities and those of low socioeconomic status are at particular risk due to barriers in healthcare. Vaccine avoidance and polarization is more widespread as trending social media is ever more virulent and effective at creating fear-mongering clickbait against vaccines. Our motivation is to support public health by quantifying the probability of an individual to get the seasonal flu vaccine and the H1N1 vaccine.

2.2 **Striving for More Adoption**

We can target the individuals who are unlikely to get vaccinated and dedicate resources to educate and facilitate access to vaccinations. Furthermore, we will explore the strongest factors influencing people’s decision to take a vaccine, although there might be a society threshold that is very difficult to break through where more investment provides marginal gain. Facilitating limited resources to communities that achieve less vaccine adoption than average could provide a large public health and social justice gain where the benefit is substantial. The common assumption is any vaccine uptake is a strong indicator for taking another vaccine and we will investigate and put that assumption to the test.

3 **Literature Review**

Exploring why people have and have not chosen to get vaccinated is a relatively common theme over the past decade. Many works focus on subsets of the population or an overall more general approach, rather than exploring the specific features and their importance in models. Examining more factors and which ones hold the most weight in determining if someone gets vaccinated will be a key focus of this article, which has not been focused in other works.

3.1 **Predicting H1N1 and Seasonal Flu:**

**Vaccine Cases using Ensemble Learning Approach**

Ayachit et al. (2020) use the same dataset from the National 2009 H1N1 Flu Survey to create nine models to predict vaccine usage (Centers for Disease Control and Prevention, 2012). Their models are reported as having high accuracy scores at predicting the actual probability of a person choosing to get vaccinated but fails to go into detail about which features were used or how the model can be interpreted for actionable insights. Interactions with the features themselves are minimal, which is a key focus of our research.

3.2  **Beliefs and Barriers Towards Flu**

**Vaccination Among College Students**

Marcell and Spurlock (2020) focus on what influences college students to get vaccinated for the seasonal flu. While college students comprise a very small portion of society, Marcell and Spurlock (2020) allege that “despite widespread exposure to a public health campaign regarding the H1N1 influenza pandemic in 2009, the majority of students still did not perceive themselves to be at risk for contracting the deadly disease” (p. 109). Their findings also describe trends in young adults and minorities in which anti-vaccination beliefs were more likely and they performed poorly on tests regarding flu facts. This work provides insight into a subset of the population, primarily that there is a lack of education regarding flu and flu-related diseases that may be contributing to the lower vaccination rates. The additional features in our research will explore how different views relate to the probability of getting vaccinated.

3.3**Social determinants of flu vaccine**

**uptake among racial/ethnic minorities in the United States**

Sanders-Jackson et al. (2021) focused on the flu vaccination rates among four major racial groups in 2017-2018 within the United States to identify differences in rates between different racial/ethnic groups. Their research found that people who identified as Latino or African American were less likely to have gotten a flu vaccine than people who identified as White or Asian. They also found that within ethnic groups, men, younger individuals, people without chronic diseases, and people who were not insured were less likely to be vaccinated, suggesting that these additional features held significance in the likelihood of people choosing to be vaccinated (Sanders-Jackson et al. 2021). These are also features that will be explored within our research, in addition to other features that may help identify specific attitudes that contribute to decisions to get vaccinated.

3.4 **Social media influencers can be used to**

**deliver positive information about the flu vaccine: findings from a multi-year study**

As a measure to improve vaccination rates, Bonnevie et al. (2021) examined how digital campaigns on social media were received by minority groups in the United States. The majority of interactions with the social media posts were positive and increased over time, which “showed that social media influencers can positively communicate health information about vaccines among a large, targeted, at-risk audience” (Bonnevie et al. 2021, p. 291). While this study supports that minorities are at high risk of not getting a flu vaccine, they were unable to measure the actual impact the social media posts had on vaccination rates. Measuring attitudes prior to the posts and after the posts would have resulted in more meaningful insights and could be an area of further research. Examining if certain attitudes have more of an effect on vaccination rates than others is something that we aim to do with the data we are using.

3.5 **Parental socioeconomic and psychological**

**determinants of the 2009 pandemic influenza A(H1N1) vaccine uptake in children**

Researching what influences parents to vaccinate their children can give some insights into the decision-making process, which was the focus of an article by Salo-Tuominen et al. (2022). They found that younger mothers and mothers with lower education were less likely to vaccinate their children from H1N1, highlighting another group that could potentially benefit from more support (2022). Interestingly, psychosocial factors such as mental illness and relationship satisfaction were not correlated with vaccination rates, showing that mental health may not be a factor, but rather societal views and general education and life experience may play a bigger role in the decision-making process to get their children vaccinated. While we will not focus on children, examining differences between age groups will likely be an area of correlation that will be addressed.

3.6 **The long road of pandemic vaccine**

**development to rollout: A systematic review on the lessons learnt from the 2009 H1N1 influenza pandemic**

Ankomah et al. (2022) highlight the importance of vaccines in the H1N1 influenza pandemic in their review of 1,056 papers related to the H1N1 pandemic vaccine. They describe the difficulties the vaccine faced as:

There was generally a low uptake of the available vaccines due to safety concerns, widespread doubts on the clinical efficacy of approved vaccines, and a general perception of a low health risk due to the moderate severity of the H1N1 influenza in 2009. (Ankomah et al., 2022, p. 739)

Their comprehensive review of the challenges the vaccine faced gives some insights into the pervasive attitude that many people do not believe they will get the diseases that vaccines protect against. Examining to which socioeconomic groups this view is common and other attitudes toward disease prevention will allow us to gain further insights into what factors into the decision to get vaccinated.

4 **Methodology**

While the data is generally clean with the exception of null values, special focus will be on handling data and determining how to generate models for each target. The main challenge is finding optimal data cleaning methods and exploring ways of handling the null values that will have the best results.

4.1 **Data acquisition and aggregation**

Data is downloaded from DrivenData (2020). Data consists of two CSV files with 26,707 rows and a respondent ID. Training data contains 35 features and training labels contain two targets (H1N1 vaccine and seasonal flu vaccine). Features include 23 numerical features, 13 of them binary, and 12 categorical features. Geographic region, employment industry, and employment occupation have been changed to strings of random characters. Data is combined on the respondent ID to create one dataframe. Two features focus on H1N1 understanding, seven focus on health-related behaviors such as hand washing, two refer to doctor recommendations, 17 explore demographics, and six contain opinions about H1N1 and seasonal flu vaccines and risks.

*4.1.1 Exploratory Data Analysis* Exploratory data analysis begins by exploring null values within the data. Five features and both targets have zero null values, while the other features range between 19 and 13,470 null values. Most contain less than 1,000 null values. The features with the most null values are employment\_occupation (13,470 nulls), employment\_industry (13,330 nulls), and health\_insurance (12,274 nulls). This represents about half of the data with no response for those features. For further data exploration, numerical null values are set to -1 and categorical values are set to ‘no\_response.’ Feature density and distribution with the targets can then be explored. H1N1 concern, knowledge, opinions about the vaccine, seasonal flu opinions, education, income, and household adults are all relatively normally distributed. Rent/own, household children, employment industry, employment occupation, race, health insurance, health worker, children under 6 months, chronic medical conditions, and healthy behaviors are all relatively skewed. Geographic location is relatively equal among categories. The H1N1 vaccine target is more heavily skewed toward ‘no’ while the seasonal flu vaccine target was close to equally distributed between ‘yes’ and ‘no’ responses. When exploring distribution toward the targets, the features that showed some correlation with someone receiving the H1N1 vaccine were the doctor recommending the H1N1 vaccine, doctor recommending the seasonal flu vaccine, being a health worker, having health insurance, believing the vaccine is effective, seeing H1N1 as high risk, employment occupation 3 or 4, and having the seasonal flu vaccine. The features with distribution showing correlation to the seasonal flu vaccine are the doctor recommending the seasonal flu vaccine, having a chronic med condition, being a health worker, having health insurance, having a higher opinion that H1N1 and the seasonal flu are high risk as well as believing the vaccines are effective, older respondents, occupation 3 or 4, and receiving the H1N1 vaccine. This is supported when exploring the correlation between features, which is performed after converting categorical features to numerical so that correlation can be calculated. Most features have less than 10% correlation to the features. The features that have correlation higher than 10% with the H1N1 vaccine target are: H1N1 concern, H1N1 knowledge, the doctor recommending the H1N1 and seasonal flu vaccines, health worker, health insurance, opinions about the risk and vaccine effectiveness for H1N1, opinions about the risk and vaccine effectiveness for the seasonal flu, and receiving the flu vaccine. The features with correlation higher than 10% with the seasonal flu target are H1N1 concern, H1N1 knowledge, washing hands, touching face, doctor recommending H1N1 vaccine, doctor recommending the seasonal flu vaccine, chronic medical condition, health worker, health insurance, opinions about the risk and vaccine effectiveness for H1N1, opinions about the risk and vaccine effectiveness for the seasonal flu, age group, race, household children, employment industry and occupation, and receiving the H1N1 vaccine. The opinion of H1N1 risk and seasonal flu risk normalized bar plots are shown in Figure 1 and Figure 2 respectively.

**Figure 1**

*Normalized Bar Plot Exploring Relationship between Opinion of H1N1 Risk and Targets*

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

*Note*. Seasonal flu vaccine is more highly correlated with opinions of H1N1 risk than H1N1 vaccine.

**Figure 2**

*Normalized bar plot exploring relationship between opinion of seasonal flu risk and targets*

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

*Note.* Correlation is similar to opinion of H1N1 risk, with slightly more correlation with the seasonal flu vaccine and less with the H1N1 vaccine.

Finally, variance is explored. Employment industry and occupation have the highest variance, as they have the most options available. Most features have low variance as they are binary, with the lowest variance for antiviral medication and face mask behavior.

4.2 **Data quality**

Data is overall clean with null values being the predominant challenge when choosing how to transform the data. Null values can be meaningful, as choosing to not answer a question provides an interesting insight into feelings. People may not be informed enough to answer a question, or they may not feel comfortable revealing some answers to the interviewer due to personal or societal biases. Data is anonymous, so it is likely that respondents are honest in their responses, but there may be some responses not entirely reflective of true feelings. Features that are not numerical contain a reasonable number of responses with 25 being the highest unique responses for employment industry and employment occupation.

*4.2.1 Data Quality Issues.* The main data quality concern is the age of the data. It is likely that some attitudes may have changed during and after the COVID-19 pandemic. The H1N1 pandemic was much smaller comparatively and had lower levels of fear and vaccination rates. The overall opinions may be similar today, but many people may have changed their views about pandemic diseases and vaccinations. Another key concern when interpreting the data is that employment occupation and employment industry are both randomized strings to represent different occupations and industries. The problem with being randomized is that the understanding of which occupations and industries relate highest to the targets is unknown. It is evident that there is correlation with two of the occupations/industries, but insights cannot be made with the information provided.

4.3 **Feature engineering**

The first step in feature engineering is converting the categorical variables to numbers using category encoding. The labels that are encoded included geo region, census, employment industry, employment occupation, employment status, rent or own, marital status, income poverty, race, education, age group, and sex. Moving into feature selection, respondent ID had all unique values and is irrelevant to keep in the modeling phase once the datasets were merged. Class imbalance is then implemented to make sure that there is enough data for each target response for training the machine learning models. One way to deal with this issue is by augmenting the minority response data through synthesizing new data from the original data, known as the Synthetic Minority Oversampling Technique (SMOTE). The first step in performing SMOTE is separating the features and targets into four different groups—each target is separated into its own training and test sets—and applying the technique on each group individually. This was done on several separated groups representing different ways to handle null values: the original data, encoded data with nulls set to -1 or no\_response depending on if the variable is numerical or categorical, nulls replaced with the median data, and the data where rows with null values are dropped. After applying the technique, the target H1N1 vaccine was successfully balanced with an even split of 21,033 rows for each response, where previously there were 21,033 ‘no’ responses and 5,674 ‘yes’ responses. This was also performed for the seasonal flu target, which was already close to balanced at 14,272 ‘no’ responses and 12,435 ‘yes’ responses, to ensure that the responses were completely balanced.

*4.3.1 Preparing train-test-validation sets.* Once the datasets are balanced, they are then split into 70-20-10 train-test-validation sets. This was again performed individually on the four different groups created earlier: the original data, encoded data, nulls replaced with median data, and nulls dropped data. This way the modeling phase will have four different datasets to test which way of handling null values is performing the best for this analysis.

*4.3.1.1 Feature scaling.* Two techniques are applied to the dataset—normalization and standardization using the scikit library. A scaler is fitted onto the training dataset, followed by transforming both training and testing sets. For the standardization technique, since the goal is to bring features to similar scale, the encoded categorical variables are either 0 or 1, hence, there is not much scale difference for applying a standardization on them. The numerical features are fifteen variables grouped under numerical columns. A group of machine learning models are then created using the normalized and standardized datasets to test their performance metrics.

4.4 **Modeling**

Several models are created to perform binary classification of both the H1N1 vaccine uptake and seasonal flu vaccine uptake. Certain models that are fed data that follow a gaussian distribution require standardization. Other models benefit from standardization where gradient descent is used as an optimization technique. Distance-based algorithms are sensitive to the feature ranges. The models used are K-Nearest Neighbors (KNN), K-means, Support Vector Machine (SVM), Linear Regression, Logistic Regression, Naïve Bayes, and Neural Network all benefit from either normalization or standardization. The ensemble models Random Forest and Adaboost consist of many weak learners and are less affected by the distribution of the dependent variables.

*4.4.1 Selection of modeling techniques.* Adaboost, Gradient Descent, and Random Forest have strong preliminary F1-scores with the default hyperparameters which is used for initial tuning. The remaining models had low F1-scores and even after exploratory tuning the improvement is minimal. Initially, a random-search of the hyperparameters was used, but progress was inconsistent. The alternate approach of using a grid-search to explore the hyperparameters is deemed the best method to systematically find the best results.

*4.4.2 Test design, i.e. training and validation datasets.* F1-scores and accuracy metrics on the test data is the primary method used to select the best models and determine the optimal datasets to use based on null value handling techniques and standardization and normalization. Confusion matrices are generated to further examine the appropriateness of the model. The models with the highest F1-scores will be used for validation metrics and final results. The final three models for each target are tuned for optimal parameters on the ideal data subset, which are then trained. The F1-score, accuracy, and AUC are then calculated for the validation dataset.

*4.4.3 Hyperparameters.* The top three models for each target are tuned to find optimal hyperparameters. The models are the same for both targets: the Random Forest, Gradient Boosting Classifier, and Adaboost. The best hyperparameters for the Random Forest models both use bootstrap=False, max\_features=`log2`, and criterion=`log loss`, while H1N1 model consisted of a max tree depth of 18 while setting the number of total trees to 120. The seasonal flu model uses a max tree depth of 19 and 90 as the number of trees. The runner-up model, Adaboost uses the following hyperparameters max\_depth=2, learning\_rate=0.2, and algorithm='SAMME.R', while the H1N1 model has n\_estimators=460 and the seasonal flu model has n\_estimators=500. It is worth noting the difference of the number of estimators for each case. The third place model, Gradient Boost, has the following set for the H1N1 model: learning\_rate=0.15, loss='log\_loss', max\_depth = 10, max\_features = 'sqrt', min\_samples\_leaf=100, min\_samples\_split=880, and subsample = 0.89. The seasonal flu model has learning\_rate=0.25, loss=`log\_loss`, max\_depth=4, max\_features=`sqrt`, min\_samples\_leaf=8, and min\_samples\_split= 1040.

5 **Results and findings**

When initially choosing the best performing models, they are evaluated based on their F1-scores. Starting with the SVM model, results showed that using the original dataset for the H1N1 virus results in the best F1-score of 0.8 while the encoded dataset performs best on the seasonal flu with an F1-score of 0.768. For Gradient Boosting the encoded data has the best metrics for the H1N1 model with an F1-score of 0.905 which is slightly lower than the dropped data model while the seasonal flu model does best with the dropped data with an F1-score of 0.791. K-Nearest Neighbors is also on the top of the list, with the H1N1 model performing best with the dropped data resulting in an F1-score of 0.818 while the seasonal flu model performed best with the median data with an F1-score of 0.725. Random Forest has decent performance with the H1N1 model performing the best with the original data, having an F1-score of 0.826 while the seasonal flu model performs the highest with the dropped data with an F1-score of 0.746. Lastly the Adaboost Classifier has the best H1N1 and seasonal flu model performances on the original data with F1-scores of 0.896 and 0.785 respectively. A graph of the area under the receiver operating characteristic curve (AUC ROC) was then created to visualize the results of the top performing models for both the H1N1 Virus and the seasonal flu. The top models are then tuned for optimal hyperparameters, and the tuned models trained and used to make predictions on the validation dataset. The Gradient Boosting Classifier model for the H1N1 virus using the nulls-dropped dataset has an F1-score of 0.903, accuracy of 0.903, and AUC of 0.905. The Adaboost H1N1 model uses the encoded-null data and has an F1-score of 0.905, accuracy of 0.905, and AUC of 0.905. The Random Forest H1N1 model with dropped null rows has an F1-score of 0.915, accuracy of 0.916, and AUC of 0.915. The Seasonal Flu Gradient Boosting Classifier model with the original data has an F1-score, accuracy, and AUC of 0.792. The Seasonal Flu Adaboost model using original data has an F1-score, accuracy, and AUC of 0.799. The final model, the seasonal flu Random Forest model using dropped null rows, has an F1-score of 0.801, accuracy of 0.801, and AUC of 0.802. These scores are shown in Table 1 with the top models for the validation data highlighted.

**Table 1**

*Model Metrics*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **F1-score** | | **Acc.** | **AUC** |
| *H1N1 Gradient Boost. Class.* | | 0.903 | 0.903 | 0.905 |
| *H1N1 Adaboost* | | 0.905 | 0.905 | 0.905 |
| *H1N1 Random Forest* | | 0.915 | 0.916 | 0.915 |
| *Flu Gradient Boost. Class.* | | 0.792 | 0.792 | 0.792 |
| *Flu Adaboost* | | 0.799 | 0.799 | 0.799 |
| *Flu Random Forest* | | 0.801 | 0.801 | 0.802 |

*Note*. The top model for both targets was the Random Forest model.

5.1**Evaluation of results**

While the Gradient Boosting Classifier models performed better on the test data, the Random Forest models performed best on previously unused validation set. This could mean that the Gradient Boosting Classifier models may have been overfitted to the training data. The H1N1 models consistently have higher scores than the seasonal flu models, most likely due to the smaller number of more highly-correlated features with the H1N1 target. The ROC AUC scores are modeled in Figure 3.

**Figure 3**

*ROC AUC Scores*

Chart

Description automatically generated

*Note.* Top three models for each target are shown.

Averaging the scores of the H1N1 and Seasonal Flu models, the best-performing Random Forest models have an average F1-score of 0.858, average accuracy of 0.859, and AUC of 0.859. The model results show that it is important to use different models for different targets, as choosing to get vaccinated is dependent on individual disease profiles. While the features used in models are similar, the top features in the H1N1 Random Forest model are doctors recommending the vaccine, the belief that the vaccine is effective, the participant having health insurance, their doctor recommending the seasonal flu vaccine, and seeing H1N1 as a risk. The seasonal flu Random Forest Model has top features of opinion that the seasonal flu vaccine is effective, belief that the seasonal flu is a risk, doctors recommending the seasonal flu vaccine, age, and region. This is shown in figure 4.

**Figure 4**

*Feature importance for H1N1 model*

*Chart, histogram

Description automatically generated*

*Note.* Top features for the H1N1 model drop off in importance after 1, 2, and 5 features.

It is clear that attitudes toward the vaccine and risk of getting the disease in question are influential on whether someone chooses to get vaccinated, which is in line with the original hypothesis.

**Figure 5**

*Feature importance for Seasonal Flu model*

**Chart, histogram

Description automatically generated**

*Note.* Top features for the seasonal flu model fall off in importance after 3, 4, and 5 features.

Region and age group are also important for people deciding whether to get the seasonal flu vaccine, showing that some areas may benefit from educational campaigns. Due to the seasonal flu being more dangerous for older individuals, there is already a higher likelihood that an older person is likely to be vaccinated, so focusing vaccine educational campaigns on younger and middle-aged people would be beneficial for improving vaccination rates. Overall attitudes and opinions are key features in predicting vaccination rates for both vaccines, which shows where information about long-term effects of diseases and risks as well as effectiveness of vaccines is important information to increase the likelihood that people will choose to get optional vaccines.

6 **Discussion**

One of the key differences in the H1N1 and seasonal flu models are the performance metrics, with the H1N1 model having a higher AUC and F1-Score at 0.915 than the seasonal flu at 0.802 and 0.801 respectively. The seasonal flu model takes in more features to make predictions at higher weights in the 0.02-0.04 range than the H1N1 model and only 13 under 0.02, while the H1N1 model has 28 features less than 0.02. While the seasonal flu vaccine had a close to even split of people who are vaccinated (12,435) versus those who are not (14,272) and the H1N1 vaccine had far fewer people vaccinated (5,674) than not (21,033), the seasonal flu model used more heavily weighted features and was less accurate than the H1N1 model. These densities are shown in figures 6 and 7.

**Figure 6**

*Density of H1N1 Vaccine Response*

*Chart, histogram

Description automatically generated*

*Note.* Target was heavily skewed to ‘no’ responses.

**Figure 7**

*Density of Seasonal Flu Vaccine Response*

*Chart, histogram

Description automatically generated*

*Note*. Target was close to evenly distributed.

Due to the differences in density, the H1N1 model was trained with a higher amount of data due to oversampling which meant that lack of variance in the training data was a concern, but the results when using the validation data showed that there was not a drop in the F1-Score or AUC for either model, suggesting that they were not overfit to the training data and they did not have a performance drop on unseen data. When exploring the features used in the models, the underlying theme of access to knowledge and information is highlighted. Doctor recommendations, opinions about vaccine efficacy, disease knowledge, and believing that the disease in question is a risk are all top factors when determining if someone is likely to be vaccinated or not. Additionally, age and location feed into the risk and education regarding diseases. One of the limitations of the data is the encoding of geographical region, which limits any insights regarding why different regions may have different attitudes toward vaccines. More questions could be used in future surveys to help target more specific attitudes and help focus on what information may help sway people to choose to be vaccinated as there are additional questions that arise out of the results, such as how often doctors are recommending vaccines and why. Additionally, with new pandemics arise additional questions and follow up data may give more timely insights. Similar to Bonnevie et al.’s (2021) findings, social media has played a bigger role in decision making and may be a key in helping to educate people about disease. It is also worthwhile to note that top features of the models for the H1N1 vaccine and the seasonal flu vaccine have different results. While both models share some similarities in top features, there are some differences, such as age being a top feature for predicting the seasonal flu vaccine outcome (0.06), but it ranks lower as a feature in predicting the H1N1 outcome (under 0.02). This shows that different vaccines have different decision processes and vaccine campaigns need to be tailored to the groups least worried about the illness. The seasonal flu vaccine is available annually and is generally seen as reliable and an important safety tool especially for young babies and children or older individuals who are more likely to be seriously ill. Vaccines like the H1N1 vaccine have minimal history due to being a response to a pandemic and the largest variables are doctor information, knowledge about the virus, opinions about vaccines, and perceived risk.

6.1 **Conclusion**

Optional vaccinations, while arguably as important as mandatory vaccinations, will continue to follow a similar trend unless they are promoted. In addition to educational background, personal beliefs, and preferences many people do not see optional vaccinations as beneficial in their lives from a medical perspective. Further investigation into multivariate impurity reduction is warranted in the transformation and selection of feature importance which may highlight overlooked risk areas. One way of promoting vaccination can be encouraging employees to get vaccinated in the workplace. Many companies coordinate annual onsite flu clinics as part of their wellness programs, which is an excellent option since many companies will cover flu vaccination through their health insurance providers. This option will also reduce employees’ time away from work and creates a healthier and more productive environment at work. By continuing to promote similar and new approaches, the percentage of people willing to get vaccinated will increase overtime. Overall, these results confirm the importance of factual information being shared with the public and the importance of trust between people and medicine. Vaccines need to be less risky than the diseases they prevent, otherwise people will choose what is perceived as the lesser risk. Providing accurate information to the public regarding health risks and ways to mitigate them for themselves and others is an important part of public health and safety, as well as finding ways to combat misinformation being spread. Providing information at a school level and beyond can help educate people to be skeptical of misinformation designed to insight fear. Ultimately, further studies are needed to find what really defines a virus as risk to a person, what defines a vaccine as trustworthy, and what doctors are recommending to their patients. Limited information regarding geographical location also leaves that variable as less informative than would be desired, and future studies into how geographical regions view vaccination importance can help direct public health campaigns. These results give further insights on where focus should be to improve vaccination rates: doctor recommendations, information about virus risks, and studies on vaccinations to show efficacy and safety. Additionally, providing more information on social media from different places that people trust, such as scientific communities and reputable sources, can help spread information in a way that will reach a wide audience. Providing facts along with examples can have a big impact on how vaccines, or not getting vaccinated, can change people’s lives. Providing information that helps people understand the greater societal impact of choices can also help improve the likelihood that people will choose to get vaccinated. Improving vaccination rates will require reliable vaccines in addition to reliable information about health concerns, risks, and easy access. Improving vaccination rates can help minimize the effect that viruses have on our communities and help improve everyone’s well-being. Using the models created, people working to increase vaccination rates can determine if someone is likely to be vaccinated or not and address factors that may be preventing people from getting vaccinated.

6.2 **Recommend next steps/future studies**

The effort to maximize the adoption of vaccines requires efficient targeting of those who are not likely to get the H1N1 or seasonal flu vaccine. Since the adoption of vaccines is a binary choice, we can modify the classification threshold from 0.5 to a higher value. This change would increase identifying the number of people who would not get the vaccine. In other words, the chance of having a Type I error, where someone is identified as taking the vaccine, but they are not vaccinated, would decrease, but the chance of a Type II error, where someone is identified as not having the vaccine, but they are vaccinated, would increase. The negative of this is that it would be a waste if efforts were directed to individuals who are already vaccinated. Weighing this extra cost with the extra benefit of identifying more vaccine-hesitant people would be a better option than not reaching out to someone who is likely to not get vaccinated. Another way of evaluating the data would be using recall because the cost of a false negative is higher than the cost of a false positive when the goal is to increase vaccination rates. Focusing on the recall, shown in formula 1, focuses on the ratio of true positives to total positive.

(1)

Using the F1-Score provided the harmonic mean and can be a strong metric for choosing models when finding a model that has a good balance between precision and recall. The formula for F1-Score is shown in formula 2.

(2)

Focusing on recall would provide a model that will have less false negatives and may be preferable depending on the cost of a vaccine campaign to ensure fewer false positives. With the COVID-19 pandemic, there is also increased information about optional vaccinations. Conducting a new survey focused on why people chose to get vaccinated or not get vaccinated could provide additional data on how trends and reasoning may have changed or stayed the same. Overall, newer information will increase the ability to predict trends and see what is impacting vaccination decisions currently and help inform how information is shared to improve vaccination knowledge and acceptance in the future.

**ACKNOWLEDGMENTS**

To our advisors and families.

**REFERENCES**

Ankomah, A. A., Moa, A., & Chughtai, A. A. (2022). The long road of pandemic vaccine development to rollout: A systematic review on the lessons learnt from the 2009 H1N1 influenza pandemic. *American Journal of Infection Control, 50*, 735-742. <https://doi.org/10.1016/j.ajic.2022.01.026>

Ayachit, S. S., Kumar, T., Deshpande, S., Sharma, N., Chauasia, K., & Dixit, M. (2020, December). Predicting H1N1 and Seasonal Flu: Vaccine Cases using Ensemble Learning approach. *2020 2nd International Conference on Advancing in Computing, Communication Control and Networking (ICACCCN),* 172-176. <https://doi.org/10.1109/ICACCCN51052.2020.9362909>

Bonnevie, E., Smith, S. M., Kummeth, C.,

Goldbarg, J., & Smyser, J. (2021, March 31).

Social media influencers can be used to

deliver positive information about the flu

vaccine: findings from a multi-year study.

*Health Education Research*, *36*(3), 286-294.

<https://academic.oup.com/her/article/36/3/28>

[6/6319647](https://academic.oup.com/her/article/36/3/28)

Centers for Disease Control and Prevention.

(2012).  *National 2009 H1N1*

*Flu Survey (NHFS)*.

<https://www.cdc.gov/nchs/nis/data_files_h1>

[n1.htm](https://www.cdc.gov/nchs/nis/data_files_h1)

Centers for Disease Control and Prevention.

(2021a).  *Quadrivalent Influenza Vaccine*.

<https://www.cdc.gov/flu/prevent/quadrivalent.>

[htm](https://www.cdc.gov/flu/prevent/quadrivalent.)

Centers for Disease Control and Prevention.

(2021b, August 26). *What are*

*the benefits of flu vaccination?*

[https://www.cdc.gov/flu/prevent/vaccine-](https://www.cdc.gov/flu/prevent/vaccine-benefits.htm)

[benefits.htm](https://www.cdc.gov/flu/prevent/vaccine-benefits.htm)

Dinerstein, C. (2018, October 26). *Influenza*

*Vaccination Is Global, But Not The Same.*

American Council on Science and Health.

Retrieved July 8, 2022, from

[https://www.acsh.org/news/2018/10/26/influ](https://www.acsh.org/news/2018/10/26/influenza-vaccination-global-not-same-12504)

[enza-vaccination-global-not-same-12504](https://www.acsh.org/news/2018/10/26/influenza-vaccination-global-not-same-12504)

DrivenData. (2020). *Flu Shot Learning: Predict*

*H1N1 and Seasonal Flu Vaccines*.

DrivenData.[https://www.drivendata.org/comp](https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/)

[etitions/66/flu-shot-learning/page/210/](https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/)

Ferdinands, J. M., Thompson, M. G., Blanton, L.,

Spencer, S., Grant, L., & Fry, A. M. (2021,

June 2). Does influenza vaccination attenuate

the severity of breakthrough infections? A

narrative review and recommendations for

further research. *Vaccine, 39*(28), 3678-3695.

<https://doi.org/10.1016/j.vaccine.2021.05.011>

Marcell, V. & Spurlock, W. R. (2020, October 1). Beliefs and Barriers Towards Flu Vaccination Among College Students. *The ABNF Journal, 31*(4), *108-112.*

Salo-Tuominen, K., Teros-Jaakkola, T., Toivonen,

L., Ollila, H., Rautava, P., Aromaa, M., Lahti, E., Junttila, N., & Peltola, V. (2022, May 4). Parental socioeconomic and psychological determinants of the 2009 pandemic influenza A(H1N1) vaccine uptake in children. *Vaccine*. <https://doi.org/10.1016/j.vaccine.2022.05.012>

Sanders-Jackson, A., Gonzalez, M., Adams, R. B.

& Rhodes, N. (2021, August 12). Social determinants of flu vaccine uptake among racial/ethnic minorities in the United States. *Preventative Medicine Reports, 24.* <https://doi.org/10.1016/j.pmedr.2021.101516>

Thompson, M. G., Pierse, N., Huang, Q. S.,

Prasad, N., Duque, J., Newbern, E. C., Baker,

M. G., Turner, N., & McArthur, C. (2018,

September 18). Influenza vaccine effectiveness

in preventing influenza-associated intensive

care admissions and attenuating severe disease

among adults in New Zealand 2012–2015.

*Vaccine, 36*(39), 5916-5925.

<https://doi.org/10.1016/j.vaccine.2018.07.028>

Tosh, P. K. (2021, May 18). *What’s the difference*

*between H1N1 flu and influenza A?* Mayo

Clinic.

<https://www.mayoclinic.org/diseases->

[conditions/swine-flu/expert-](https://www.mayoclinic.org/diseases-)

[answers/influenza-a/faq-20058309](https://www.mayoclinic.org/diseases-)