**Predicting Vaccine Uptake**

Akash Yadav (ay1652@nyu.edu)

Aparna Bhutani (ab8473@nyu.edu)

Tatsuya Haga (th2254@nyu.edu)

**Purpose of the Project**

The pandemic of the novel H1N1 influenza virus, also known as the swine flu pandemic, occurred in March 2009 and lasted till August 2010. The usual flu vaccine did not provide protection against this virus and the vaccine for it was released in late October 2009. Until the discovery of the vaccine, the virus spread across major parts of the world such as the United States, Mexico, and Spain, with the number of cases around the world being approximately 1.6 million. The total number of deaths all over the world was a staggering number between 150,000 to 575,000. Taking a vaccine early helps in reducing the severe effects of the virus and improving the health of the individual infected.

Here in our work, our goal is to use machine learning techniques to predict the vaccine uptake for the H1N1 influenza virus considering different variables provided in our dataset. We also look at the key factors responsible for predicting the vaccine uptake for H1N1. The adequate rates of vaccine uptake play a critical role in containing the spread of a virus and its effects which is crucial for the health of people and the proper functioning of urban society. This project helps in finding the factors which majorly influence why a person will/ will not take the vaccine so that we can focus specifically on people who won't.

Ensuring enough people take the vaccine will be an imminent task for governments around the world once the vaccine for the COVID-19 becomes available. While we acknowledge that there are differences like the enforcement of lockdown measures, we believe that the H1N1 influenza virus outbreak shares important commonalities with the COVID-19, particularly the novelty of the virus and how the vaccine became available in midst of the outbreak. Therefore, we believe that the results obtained from our project will be useful for the governments and doctors to achieve a high rate of uptake of the prospective COVID-19 vaccine. Based on our results, we present several recommendations for the vaccination campaign for COVID-19 once it becomes available.

**Description of Related Work**

There have been several studies that analyzed the factors that affected the uptake of the H1N1 influenza vaccine. Among them, two studies have targeted the general U.S. population and considered comprehensive factors. Kumar et al. (2012) utilized the social-ecological model to analyze the determinants of vaccine uptake, finding that attitudes toward the virus, the social influence of family and friends, information from the health care provider, risk of the virus in the community, and access to healthcare all affected the vaccine uptake.[[1]](#footnote-1) Galarce, Minsky, and Viswanath (2011) analyzed how the socioeconomic status, demographics, and personal beliefs affect vaccine uptake.[[2]](#footnote-2) They found that age, urbanity, belief in the safety of the vaccine, and seasonal flu vaccine uptake had strong associations with the vaccine uptake. Other studies have focused on certain subgroups, such as school teachers or healthcare workers.[[3]](#footnote-3)[[4]](#footnote-4)

Our approach differs from the previous studies mainly in two ways. First, our dataset contains a wide range of variables that have not been explored in the previous studies. These include various behavioral factors, such as reducing social contacts or conducting hygiene practices, attitudes toward the virus and the vaccine, and socioeconomic status such as family size. Next, the previous studies primarily use logistic regression to analyze the factors, while we attempt to apply different machine learning methods. Given the decision boundary may not necessarily be linear, as logistic regression assumes, our model could have better prediction accuracy.

**Data**

**Data Source**

We used the dataset from the National 2009 H1N1 Flu Survey by the U.S. Department of Health and Human Services, available at DrivenData.[[5]](#footnote-5) The survey targeted all persons in the U.S. older than six months. It took place between October 2009 and June 2010, inquiring about vaccination status, influenza-related behaviors, opinions toward vaccination, recent medical history, socioeconomic status, and demographic characteristics.[[6]](#footnote-6)

**Sample Size**

The sample size is N=26,706.

**Variables**

The variables in the dataset are as below.

|  |  |  |
| --- | --- | --- |
| Variable Name | Definition | Coding |
| *h1n1\_vaccine* | Has taken H1N1 flu vaccination | 0=No; 1=Yes |
| *seasonal\_vaccine* | Has taken seasonal flu vaccination | 0=No; 1=Yes |
| *h1n1\_concern* | Level of concern about the H1N1 flu | 0=Not at all concerned; 1=Not very concerned; 2=Somewhat concerned; 3=Very concerned |
| *h1n1\_knowledge* | Level of knowledge about H1N1 flu | 0=No knowledge; 1=A little knowledge; 2=A lot of knowledge |
| *behavioral\_antiviral\_meds* | Has taken antiviral medications | 0=No; 1=Yes |
| *behavioral\_avoidance* | Has avoided close contact with others with flu-like symptoms | 0=No; 1=Yes |
| *behavioral\_face\_mask* | Has bought a face mask | 0=No; 1=Yes |
| *behavioral\_wash\_hands* | Has frequently washed hands or used hand sanitizer | 0=No; 1=Yes |
| *behavioral\_large\_gatherings* | Has reduced time at large gatherings | 0=No; 1=Yes |
| *behavioral\_outside\_home* | Has reduced contact with people outside of own household | 0=No; 1=Yes |
| *behavioral\_touch\_face* | Has avoided touching eyes, nose, or mouth | 0=No; 1=Yes |
| *doctor\_recc\_h1n1* | H1N1 flu vaccine was recommended by doctor | 0=No; 1=Yes |
| *doctor\_recc\_seasonal* | Seasonal flu vaccine was recommended by doctor | 0=No; 1=Yes |
| *chronic\_med\_condition* | Has any of the following chronic medical conditions: asthma or an other lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines taken for a chronic illness. | 0=No; 1=Yes |
| *child\_under\_6\_months* | Has regular close contact with a child under the age of six months | 0=No; 1=Yes |
| *health\_worker* | Is a healthcare worker | 0=No; 1=Yes |
| *health\_insurance* | Has health insurance | 0=No; 1=Yes |
| *opinion\_h1n1\_vacc\_effective* | Respondent's opinion about H1N1 vaccine effectiveness | 1=Not at all effective; 2=Not very effective; 3=Don't know; 4=Somewhat effective; 5=Very effective |
| *opinion\_h1n1\_risk* | Respondent's opinion about risk of getting sick with H1N1 flu without vaccine | 1=Very Low; 2=Somewhat low; 3=Don't know; 4=Somewhat high; 5=Very high |
| *opinion\_h1n1\_sick\_frm\_vacc* | Respondent's worry of getting sick from taking H1N1 vaccine | 1=Not at all worried; 2=Not very worried; 3=Don't know; 4=Somewhat worried; 5=Very worried |
| *opinion\_seas\_vacc\_effective* | Respondent's opinion about seasonal flu vaccine effectiveness | 1=Not at all effective; 2=Not very effective; 3=Don't know; 4=Somewhat effective; 5=Very effective |
| *opinion\_seas\_risk* | Respondent's opinion about risk of getting sick with seasonal flu without vaccine | 1=Very Low; 2=Somewhat low; 3=Don't know; 4=Somewhat high; 5=Very high |
| *opinion\_seas\_sick\_from\_vacc* | Respondent's worry of getting sick from taking seasonal flu vaccine | 1=Not at all worried; 2=Not very worried; 3=Don't know; 4=Somewhat worried; 5=Very worried |
| *age\_group* | Age group of respondent | 65+ years; 55-64 years; 45-54 years; 35-44 years; 18-34 years; |
| *education* | Self-reported education level | College graduate; Some College; 12 Years;<12 Years |
| *race* | Race of respondent | White; Black; Hispanic; Other or multiple |
| *sex* | Sex of respondent | Female; Male |
| *income\_poverty* | Household annual income of respondent with respect to 2008 Census poverty thresholds | <=$75,000, above poverty; >$75,000; Below poverty |
| *marital\_status* | Marital status of respondent | Married; Not married |
| *rent\_or\_own* | Housing situation of respondent | Own; Rent |
| *employment\_status* | Employment status of respondent | Employed; Not in labor force; Unemployed |
| *hhs\_geo\_region* | Respondent's residence using a 10-region geographic classification defined by HHS | Coded as random character strings |
| *census\_msa* | Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census | MSA, not principal city; MSA, principal city; Non-MSA |
| *household\_adults* | Number of other adults in household | Continuous |
| *household\_children* | Number of children in household | Continuous |
| *employment\_industry* | Type of industry respondent is employed in | Coded as random character strings |
| *employment\_occupation* | Type of occupation of respondent | Coded as random character strings |

**Preprocessing**

The dataset that we have used contains categorical variables as shown in the table above. In the exploratory phases, it was observed that the data consisted of a lot of missing values. Dropping them would have rendered only half of the dataset useful, therefore missing value imputation became a necessary procedure of our analysis method. Since our data represented categorical values, we had to figure a way to impute missing values using techniques other than the Simple Imputation methods provided by the sklearn package since most variables we were dealing with either represented an ordinal scaling or were binary. After going through techniques for categorical variable imputation, we came across two methods that would have helped our specific case, namely KNN Imputation and Multiple Imputation by Chained Methods (MICE).

After ordinally encoding the values of our dataset, we ran both the imputation techniques and were able to fit the predicted variables from each method for our multirow missing values use case. We used approximations to assign back the original categories of the data. We eliminated employment-related variables (*employment\_industry* and *employment\_occupation)* since the imputation required us to predict the values for more than half of the dataset.

**Methods**

To create a prediction model for the uptake of the H1N1 vaccine, we primarily used the Decision Tree method provided in the sklearn package. The target attribute of the model is the uptake of the vaccine (*h1n1\_vaccine*). We used the GridSearch method to determine the optimal parameter for the depth of the tree, using the training dataset to build the models and testing dataset to compare the AUC score. To help us understand which predictors had more influence, we calculated the Gini importance for each variable. Since the Decision Tree in the sklearn package does not handle categorical values, we transformed all nominal variables to dummy variables prior to the analysis.

In addition, given the instability of the Decision Tree method,[[7]](#footnote-7) we also used the RandomForest method to create a similar model to check the robustness of our result. After optimizing the depth of the trees using the GridSearch method, we constructed the RandomForest model and calculated the Gini importance of the variables. We then compared them with the Gini importance from the Decision Tree model.

**Result**

**1. MICE Imputation Model**

First, we present the results from the model using the dataset that was imputed by the MICE Imputation method.

Using the GridSearch method, we determined that the optimal depth for the Decision Tree was 5. The tree generated using the optimized parameter is shown in Appendix A. The out-sample AUC score of the model was 0.822. The result of the feature importance calculation of the variables is shown below. Any variables with feature importance of zero are omitted from the chart.



From the tree and the feature importance, we found that getting a recommendation from the doctor on taking the vaccine was the strongest predictor. The variable appeared at the top of the tree, and most of those who were predicted to take the vaccine had a recommendation from the doctor. Also, opinion about the effectiveness of the vaccine had a strong influence on the prediction as well. Specifically, those who believed that the vaccine was either somewhat effective or very effective were more likely to be predicted to take the vaccine. People who believed the vaccine was not at all effective or not very effective were more likely to be predicted as not to take the vaccine. Lastly, the opinion about the risk of getting the H1N1 virus was also a key factor. For those who did not get a recommendation from the doctor, only those who had thought that the infection risk was high, thought that the vaccine was effective, and was a healthcare worker were predicted to take the vaccine.

Such a result was confirmed by the RandomForest method as well. The out-sample AUC score of the RandomForest model was 0.836, and the feature importance for the variables in the model are shown below. We found that doctor recommendation, opinion about the risk of the virus, and the opinion about the vaccine's effectiveness had the highest importance, similar to the Decision Tree model.

****

**2. KNN Imputation Model**

Next, we present the results from the model using the dataset that was imputed by the KNN Imputation method.

Using the GridSearch method, we determined that the optimal depth for the Decision Tree was 6. The tree generated using the optimized parameter is shown in Appendix B. The out-sample AUC score of the model was 0.822. The feature importance for the variables are as below. Any variables with feature importance of zero are omitted from the chart.



From the tree and the feature importance, we find similar results with the MICE Imputation model. The doctor's recommendation was the most important predictor, appearing at the top of the tree. Opinion about the effectiveness of the vaccine and opinion about the risk of getting the virus were also the second and the third most important variables. Being a healthcare worker was also more likely to be predicted as taking the vaccine, in line with our result from the MICE Imputation model.

We also see that demographics, such as race, influence the prediction in this model. However, in the RandomForest model, the demographics had no importance, indicating it could have been due to the instability of the Decision Tree method. The out-sample AUC score of the RandomForest model was 0.835, and the feature importance for the variables in the model are shown below. We see that demographics were not important for predicting the vaccine uptake.



**3. Comparison with the Seasonal Flu Vaccine Uptake**

The results are in line with the general behavioral attitudes of people during a pandemic. One of the central emotional responses during a pandemic is fear.[[8]](#footnote-8) Though people have different perceptions about different vaccines, a recommendation from their doctor works well to comfort them during such a time. Risk perceptions are also amplified during such troubled times. In order to distinguish the motivators of vaccine uptake during the pandemic, we compared the results of our model by doing a separate analysis where we used uptake of the seasonal flu vaccine as the target variable. The important features that we observed from this analysis differed quite differently from the results for the uptake of the H1N1 vaccine.

In the models using the MICE Imputation, opinions were seen to matter more than doctor’s recommendations in the case of the seasonal flu vaccine. Demographic features such as age, income, education, and race also turned up to be important factors in determining the vaccine uptake. Older populations with income above $75,000 were more prone to take the vaccine whereas younger adults without a college degree were seen to not take the vaccine. Similar results were seen when using the Decision Tree method with KNN imputation for missing values. Chronic medical conditions along with the age of the individual also showed up as an important factor determining the individual's propensity.

Random forests for both imputation methods also saw most of these parameters repeat making a strong case for demographic factors determining the seasonal flu vaccine uptake. The accuracy scores that we were able to observe for all our methods were in the range of 80-85% for all models.

The results show that factors that affect the uptake of the vaccine were different for the H1N1 vaccine and seasonal flu vaccine. We found that while demographic features mattered for the seasonal flu vaccine uptake, doctor recommendation, risk perceptions, and opinions about the vaccine effectiveness mostly mattered for the pandemic H1N1 vaccine uptake.

**Conclusion**

Using the Decision Tree and Random Forest methods, we were able to build a high-accuracy prediction model of the H1N1 vaccine uptake. We found that the doctor's recommendation, opinion on the risk of getting the virus, and opinion on the effectiveness of the vaccine were the key predictor for whether a person took the vaccine or not.

While the H1N1 virus and the COVID-19 are different, we believe the result provides us important lessons for the expected vaccine for the COVID-19. Based on the results, we present several recommendations regarding the vaccination campaign for the COVID-19 once it becomes available.

* **Mobilize doctors to recommend people to take the vaccine**.   
  Doctor recommendation was the most important factor influencing whether a person took the H1N1 vaccine or not. We recommend that once the vaccine for the COVID-19 becomes available, the governments cooperate with doctors to ensure that people are given recommendation to take the vaccine from doctors whenever possible.
* **Communicate the risk of getting the COVID-19 virus and the effectiveness of the vaccine.**Opinions on the risk of getting the H1N1 virus without the vaccine and on whether the vaccine was effective or not were important factors influencing the H1N1 vaccine uptake. While people may be much more aware of infection risk of the COVID-19 than they were of the H1N1 virus, the government needs to actively communicate to the people the effectiveness of the COVID-19 vaccine once it becomes available.

1. Kumar, S., et al. (2012). The Social Ecological Model as a Framework for Determinants of 2009 H1N1 Influenza Vaccine Uptake in the United States. *Health Education & Behavior, 39*(2):229–243. doi:[10.1177/1090198111415105](https://doi.org/10.1177/1090198111415105) [↑](#footnote-ref-1)
2. Galarce, E.M., Minsky, S., & Viswanath, K. (2011). Socioeconomic status, demographics, beliefs and A(H1N1) vaccine uptake in the United States. *Vaccine 2011, 2*9(32):5284–5289. doi:[10.1016/j.vaccine.2011.05.014](https://doi.org/10.1016/j.vaccine.2011.05.014) [↑](#footnote-ref-2)
3. Gargano, L.M., et al. (2011). Seasonal and 2009 H1N1 influenza vaccine uptake, predictors of vaccination and self-reported barriers to vaccination among secondary school teachers and staff. *Human Vaccines, 7*(1):89-95. doi:[10.1177/1090198111415105](https://doi.org/10.1177/1090198111415105) [↑](#footnote-ref-3)
4. Henriksen Hellyer, J.M., et al. (2011). Attitudes toward and Uptake of H1N1 Vaccine among Health Care Workers during the 2009 H1N1 Pandemic. *PLoS One 2011, 6*(12): e29478. doi:[10.1371/journal.pone.0029478](https://dx.doi.org/10.1371%2Fjournal.pone.0029478) [↑](#footnote-ref-4)
5. DrivenData. (n.d.). Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines. https://www.drivendata.org/competitions/66/flu-shot-learning/ [↑](#footnote-ref-5)
6. Centers for Disease Control and Prevention National Center for Immunization and Respiratory Diseases and National Center for Health Statistics. (2012). National 2009 H1N1 Flu Survey (NHFS) A User’s Guide for the Public-Use Data File. ftp://ftp.cdc.gov/pub/Health\_Statistics/NCHS/Dataset\_Documentation/NIS/nhfs/nhfspuf\_DUG.PDF [↑](#footnote-ref-6)
7. Li, R.H. & Belford, G.G. (2002). Instability of Decision Tree Classification Algorithms. [*KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*](https://dl.acm.org/doi/proceedings/10.1145/775047)*, July 2002*:570–575. doi:[10.1145/775047.775131](https://doi.org/10.1145/775047.775131) [↑](#footnote-ref-7)
8. Bavel, J.J.V., Baicker, K., Boggio, P.S. et al. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nat Hum Behav*. doi:10.1038/s41562-020-0884-z [↑](#footnote-ref-8)