Flu Shot Learning

Team: I Love Data
Members: Els Dai, Jieru Shen, Yanjun Wan, Crystal Wu, Flora Zhang

Background:

Coronavirus disease, also known as COVID-19, is undoubtedly one of the most hot-button global issues in 2020. With almost 70 million people around the world infected and more than 1.5 million deaths caused by COVID-19 are reported, tens of millions of people are still at risk of COVID-19 infection. This COVID-19 pandemic will not halt unless a COVID vaccine, which is possibly the only exit strategy for the pandemic, is available to the world. As companies such as Pfizer and Moderna both showed that their vaccines are highly effective and more vaccines are continuing to be developed, the question now is who should we prioritize to receive the vaccines? But what caught our attention is the question that what groups of people are more likely to receive vaccines and believe that vaccines can be effective in preventing infections?

Introduction:

For this project, we utilized a data file downloaded from Kaggle which is a dataset of the National 2009 H1N1 Flu survey provided by the National Center for Health Statistics. The dataset contains key background information including the social and economic background of each respondent, his or her personal opinions and knowledge on H1N1 and seasonal flu, and whether the respondent received the H1N1 or seasonal flu vaccines. Our goal is to construct a machine learning model to predict whether an individual will receive vaccines and identify key characteristics or features that will more likely lead individuals to receive vaccines. This prediction model can be a guidance on whether a person will receive COVID-19 vaccines to further answer the question on who will possibly receive the vaccines first.

The dataset consists of 35 predictor variables and 2 dependent variables (Appendix 1). Around a quarter of the sample had received H1N1 vaccines and around half of the sample had received seasonal flu vaccines (Appendix 2). We used different machine learning algorithms such as logistic regression, ensemble learning, and KNN to construct prediction models and make comparisons to select the model with the highest accuracy. Lastly, we identified features that are important in determining whether a respondent received H1N1 or seasonal flu vaccines or not.

Data Preprocessing:

Step 1: Since the data file with all the features of respondents and the data file with whether respondents had received H1N1 or seasonal flu vaccines are separated, we merged the two files into one data frame based on the same respondent ID.

Step 2: We created two data frames based on the original data frame. The first data frame we created, df_h1n1, is a data frame that is used to predict whether an individual had received H1N1 vaccines. The second data frame, df_seas, is a data frame that is used to predict whether an individual had received seasonal flu vaccines.

Step 3:

a) Numerical variables:
We examined the distribution of each numerical variable and filled

null values with the median of each variable to more accurately represent these features (Appendix 3). We used the median for two reasons. First, the distribution of each feature is not normally distributed, so median could better represent the data. Second, the respondents' answers are discrete. For example, they are asked to rate their attitude towards the risks of h1n1 from 1-5. Using the mean such as a float number is inappropriate here.

- H1n1_concern
- H1n1_knowledge
- Opinion_h1n1_vacc_effective
- Opinion_h1n1_risk
- Opinion_h1n1_sick_from_vacc
- Opinion_seas_vacc_effective
- Opinion_seas_risk
- Opinion_seas_sick_from_vacc

b) Categorical variables:
We converted categorical data to binary variables using one-hot encoding methods to place equal importance on each category in each variable.

- Age_group
- Education
- Race
- Income_poverty
- Sex
- Marital_status
- Rent_or_own
- Employment_status
- Census_msa

Step 4: We drop columns that provide uninterpretable information including employment_occupation, employment_industry, and hhs_geo_region. We also dropped health_insurance because it only has half of the data filled. Besides, h1n1 and seasonal flu shots are free. Whether having health insurance or not does not play an important role here.

Step 5: Last step before we began the training of prediction models, we split both datasets into training datasets and testing datasets using the train_test_split function with testing datasets size as 0.2.

Model Training and Comparison:

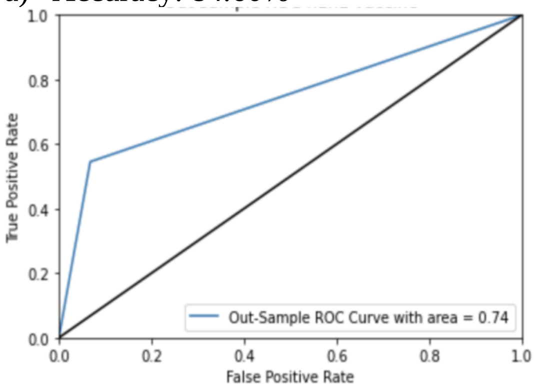
We split the data with 80% training and 20% testing and fit various models to find out which has the most predictive power.

In our preliminary analysis, we examined the linear regression model, the logistic regression model, random forest algorithm, Bootstrapping and the KNN model. We then remove the linear regression model and Bootstrapping algorithm that have lower accuracy when we compare the predicted result to the actual results in the test set.

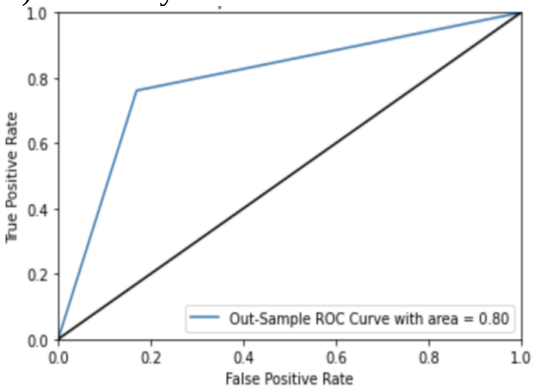
1) **Logistic regression model**
Logistic regression model is a classification algorithm that is widely used for predicting a categorical dependent variable value. In our case, 0 or 1 respectively represents whether an individual had received vaccines or not.

To avoid the problem of overfitting, we removed variables with low p-values and used the remaining variables to fit the new logistic regression model. We used accuracy to denote the R-squared value which indicates the goodness of fit. The accuracy for the H1N1 prediction and the seasonal flu prediction and the ROC curves are as follows:

a) Accuracy: 84.60%



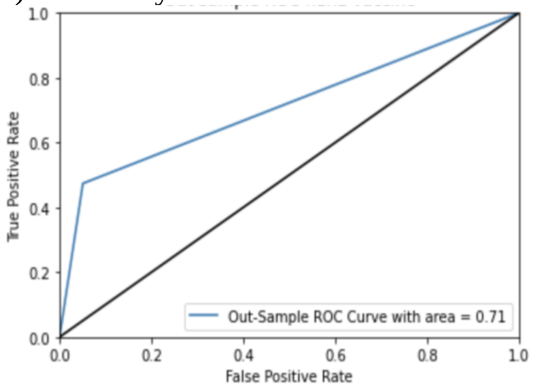
b) Accuracy: 79.75%



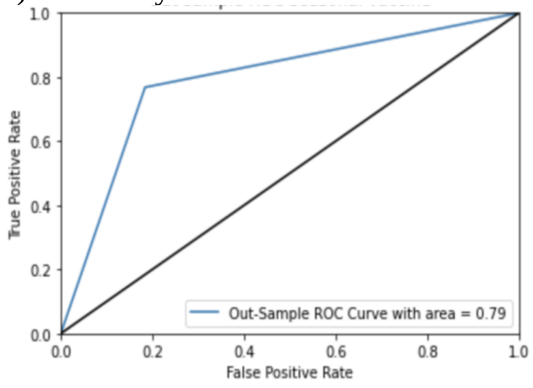
2) **Random Forest**

We also used machine learning to help predict whether a person will take the flu vaccine or not. The first algorithm that we use is Random Forest. It consists of many individual decision trees that can be used for classification tasks. It adds randomness to the model by splitting on a random subset of features to obtain accurate predictions. The accuracy and the ROC curves are as follows:

a) Accuracy: 84.31%



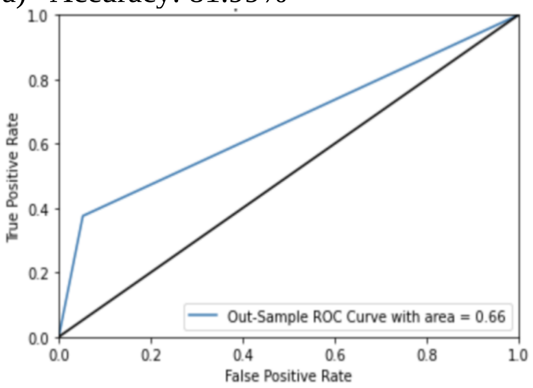
b) Accuracy: 80.40%



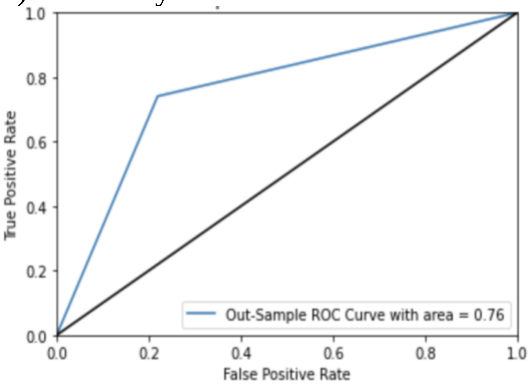
3) **K-Nearest Neighbors (KNN)**

K Nearest Neighbor (KNN) Algorithm calculates the distance between one point and every other point, and uses the top K nearest neighbors to classify the target point. Basically, it examines the features of a new data point and determines how similarly it resembles a data point in the training data set. In this case, we predicted whether each respondent in the testing data set will receive vaccines based on the feature similarities and the number of neighbors that we use is 33. We got this number by running the model recursively with different numbers of neighbors and found the number associated with the lowest error rate. The accuracy for both models and the ROC curves are as follows:

a) Accuracy: 81.99%



b) Accuracy: 60.73%



Conclusions:

From comparing the above three models we constructed, we came to a conclusion that the logistic regression models for H1N1 and seasonal flu are both better at predicting whether a respondent had received the vaccines or not with comparatively higher accuracy score and AUC.

By examining the result of the logistic regression model, we find that the variables that have higher influences on the results for the flu vaccine include: doctor's recommendation, whether the seasonal flu vaccine is taken, whether the person is a health care worker, the opinion on the risk of the flu and the effectiveness of the vaccine, and whether the person has contact with a susceptible population.

Among these predictors, the doctor's vaccine recommendation and past H1N1 or seasonal flu vaccine record are more influential than other predictors. However, it is surprising that the doctor's recommendation to take the seasonal flu shot will decrease the probability for a person to take the H1N1 flu shot, and vice versa.

Further Discussion:

Limitations and further improvements:

1. The sample size we obtained from the data file is around 20,000 data. If we want to improve the accuracy of our models and generalize our findings, we expect larger sample size and more statistics on different diseases.
2. Some of the opinion-oriented input features that we extracted from the dataset have ambiguous criteria for ratings, such as the level of the concern for the disease. If these features can be quantified in a more explanatory reasoning, we can potentially perform better data processing methods.
3. According to some important features that we identified, whether the person is a healthcare worker is also one of the determining features in the models. Therefore, if we are able to acquire more information regarding respondents' occupations and some other potential critical features such as how often the respondents visit their doctors or respondents' current locations, we believe our models can also be improved with more input features.
4. We can improve our current models by taking the correlations between input features into consideration when selecting variables (Appendix 4).
5. The machine learning algorithms we used are regression models and ensemble learning that are covered in class, but in the future, we can also try a gradient boosting framework like Light Gradient Boosting Machine and compare the performance to the accuracy score of current models we constructed.

Application:

The logistic regression model that we constructed does have a certain extent of power in predicting whether a person, given social, demographic, and economic background, and opinions on infectious disease, will receive vaccines. With this model, we can apply it to the current situation of COVID-19 pandemics to predict what groups of people are more likely to receive the vaccine. Based on the above analysis, some significant characteristics we identified that are critical in predicting who will receive vaccines are doctors' recommendations to receive the vaccine, whether an individual had received vaccines for other infectious disease, and an individual's opinions on the infectious disease. Hence, if we aim to have a powerful model to predict who will be more likely to receive the vaccine for COVID-19, it is practical to have a survey that assesses an individual's opinions on COVID-19. In this case, these opinion-oriented features can possibly add value and predict power to the flu shot learning model for COVID-19.

Appendix

Appendix 1:

List of features:

For all binary variables: 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. (binary)
- behavioral_avoidance - Has avoided close contact with others with flu-like symptoms. (binary)
- behavioral_face_mask - Has bought a face mask. (binary)
- behavioral_wash_hands - Has frequently washed hands or used hand sanitizer. (binary)
- behavioral_large_gatherings - Has reduced time at large gatherings. (binary)
- behavioral_outside_home - Has reduced contact with people outside of own household. (binary)
- behavioral_touch_face - Has avoided touching eyes, nose, or mouth. (binary)
- doctor_recc_h1n1 - H1N1 flu vaccine was recommended by doctor. (binary)
- doctor_recc_seasonal - Seasonal flu vaccine was recommended by doctor. (binary)
- chronic_med_condition - Has any of the following chronic medical conditions: asthma or another 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. (binary)
- child_under_6_months - Has regular close contact with a child under the age of six months. (binary)
- health_worker - Is a healthcare worker. (binary)
- health_insurance - Has health insurance. (binary)
- 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_from_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.
- education - Self-reported education level.
- race - Race of respondent.
- sex - Sex of respondent.
- income_poverty - Household annual income of respondent with respect to 2008 Census poverty thresholds.
- marital_status - Marital status of respondent.
- rent_or_own - Housing situation of respondent.
- employment_status - Employment status of respondent.
- hhs_geo_region - Respondent's residence using a 10-region geographic classification defined by the U.S. Dept. of Health and Human Services. Values are represented as short random character strings.
- census_msa - Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.
- household_adults - Number of other adults in household, top-coded to 3.
- household_children - Number of children in household, top-coded to 3.
- employment_industry - Type of industry respondent is employed in. Values are represented as short random character strings.
- employment_occupation - Type of occupation of respondent. Values are represented as short random character strings.

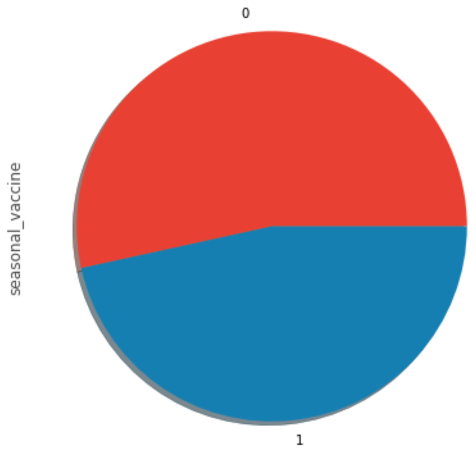
List of Target Variables:
Both are binary variables: 0 = No; 1 = Yes

- h1n1_vaccine - Whether respondent received H1N1 flu vaccine.
- seasonal_vaccine - Whether respondent received seasonal flu vaccine.

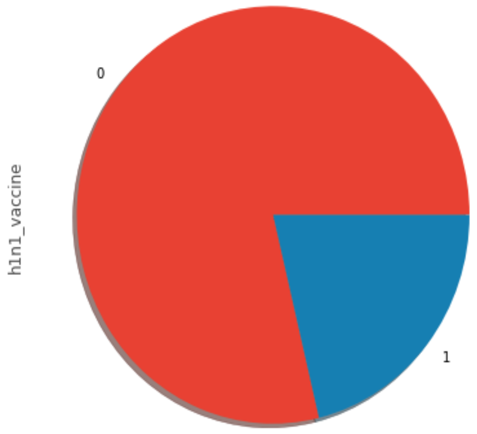
Appendix 2:

Percentage of people that have received the season flu vaccine and the H1N1 flu vaccine

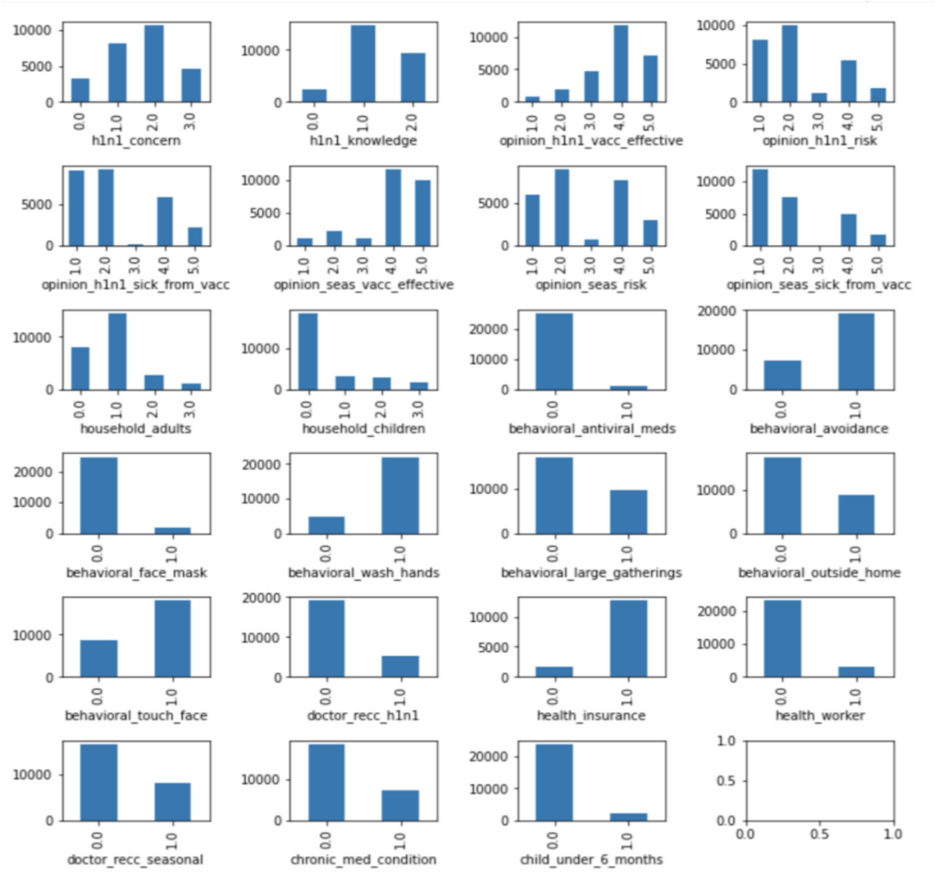
About half of people received the seasonal flu vaccine



About Less than 1/4 of people received the H1N1 flu vaccine

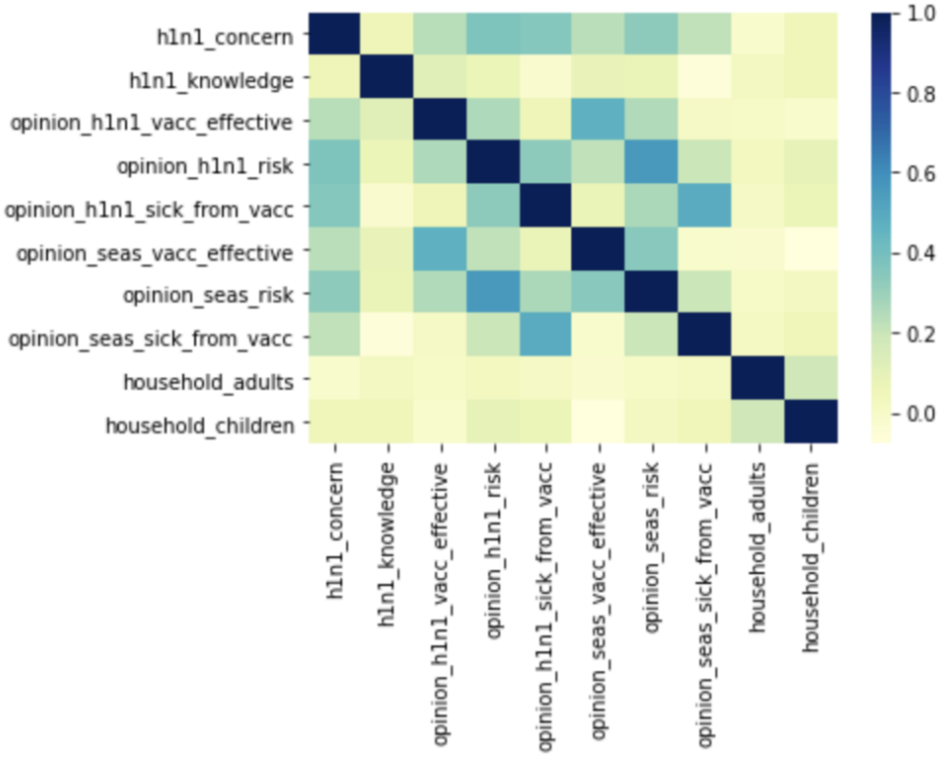


Appendix 3:
Distribution of responses for each numerical predictor variable:



Appendix 4:

Heatmap for selected important predictor variables:


h1n1_concern
h1n1_knowledge
opinion_h1n1_vacc_effective
opinion_h1n1_risk
opinion_h1n1_sick_from_vacc
opinion_seas_vacc_effective
opinion_seas_risk
opinion_seas_sick_from_vacc
household_adults
household_children