Introduction

The goal of this analysis is to identify the factors that have the greatest impact on individuals' ability or decision to take the H1N1 and seasonal flu vaccine. The result of this analysis can help the respective government agencies to come up with better ways when addressing the concerns of those communities that are unable or reluctant to take the vaccine.

Data Source

The dataset used in this analysis comes from the United States National Center for Health Statistics.[1] Each row in the dataset represents one person who responded to National 2009 H1N1 Flu Survey

A brief description of the dataset is as follows:

This data set contains 26,707 rows. The two target variables are: H1N1 vaccine and seasonal flu vaccine. There are 35 total features that has been gathered in this survey and the analysis will determine which features have the highest impact on whether someone will get vaccinated or not.

Features Description[2]

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 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. (binary)

[1] The reference for this project is: Drivendata.org

[2] This section features description refrence is: Drivendata.org

- 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 the 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.

Exploring The Data

In the first step, Figure 1 shows imbalance of the classes for H1N1 flu vaccine vs. the balanced class distribution of the seasonal flu vaccine. That means majority of the respondents haven't received the H1N1 vaccine. Processing steps should be applied when creating a predictive model for H1N1 vaccine.

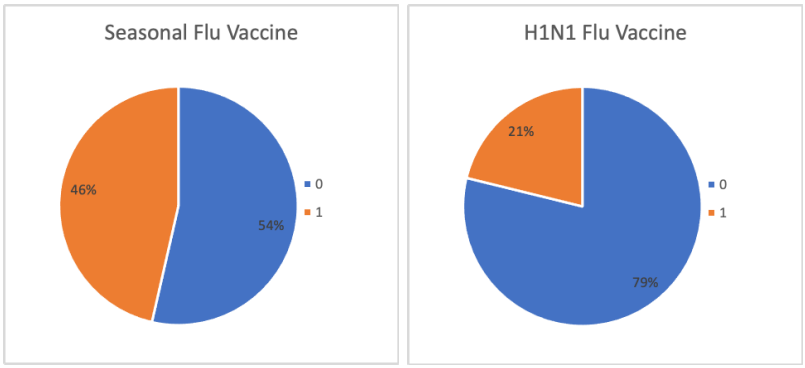

Figure 1- Target variables class distribution

Let's also evaluate the correlation between these two target variables. From Figure 2 below, the Phi coefficient has been calculated to be 0.377 which indicates slightly positive correlation between getting H1N1 vaccine and Seasonal vaccine:

Count of respondent_id	Column Labels		
H1N1 \ Seasonal	0	1	Grand Total
0	13295	7738	21033
1	977	4697	5674
Grand Total	**14272**	**12435**	**26707**

Figure 2- H1N1 and seasonal vaccine cross-table

Phi Coefficient: $\emptyset = \frac{13{,}295 \text{ x } 4{,}697 - 7{,}738 \text{ x } 977}{\sqrt{14{,}272x21{,}033\ x12{,}435x5{,}674}} = 0.377$

Next, we will investigate the distribution of some of the features within the dataset and categorize them into three groups:

1- Features with relatively balanced class:

Figure 3- Balanced class features

2- Feature with relatively imbalanced class:

- 79% of the respondents are identified as white. Pre-processing steps are needed for this feature before running a data model. For the purpose of this study this feature is disregarded.

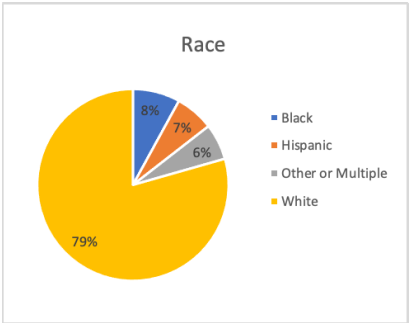

Figure 4- Imbalanced feature

3- Features with high blank percentage response:

- 17% and 46% of respondents respectively left the income and health insurance question blank.

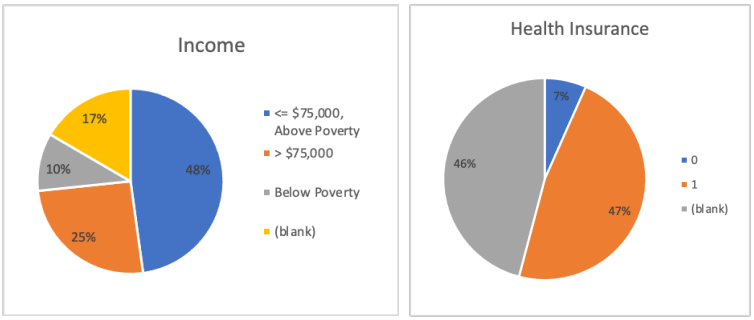

Figure 5- Features with high blank response percentage

- Nearly 50% of the respondents left employment occupation and industry blank. This might be a survey related issue that 23 employment occupation and 21 employment industry might not be inclusive enough to capture many of respondents' occupations which needs further investigation. For the purposes of this study, we disregard these two features due to many blank instances.

Methodology:

The proportion of the respondents who received the vaccine to the ones who didn't when group by each feature is calculated as variable P.

$$P = \frac{Count\ of\ respondents\ who\ recieved\ the\ vaccine}{Count\ of\ respondents\ who\ didn't\ recieve\ the\ vaccine}$$

H1N1 Flu Vaccine:

For features like H1N1 Concern which have multiple levels to choose from, the variance of P from one level to the other (from 0 to 3: 0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned.) shows how significant that feature is (H1N1 Concern) in receiving the vaccine. See Table 1 for line coefficient and correlation of these features with P for H1N1 vaccine. This table is ordered by features with highest correlation.

Feature	Line Coefficient	Correlation
H1N1 Concern	0.07	0.98
Opinion Seasonal Risk	0.14	0.97
H1N1 Knowledge	0.11	0.96
Income Poverty	0.05	0.95
Census MSA	0.01	0.94
Opinion H1N1 Risk	0.23	0.93
Opinion Seasonal Vacc Effectiveness	0.08	0.92
Employment Status	0.04	0.88
Opinion H1N1 Vacc Effectiveness	0.14	0.85
Age Group	0.02	0.83
Education	0.03	0.66
Opinion H1N1 Sick from Vacc	0.04	0.55
Opinion Seasonal Sick from Vacc	0.01	0.17
Household Adults	-0.01	-0.51
Household Children	-0.01	-0.55

Table 1- Multiple choices features correlation to P for H1N1 vaccine

As an example, to calculate the values of the first row for Table 1, a linear regression model is created between H1N1 Concern Level and P (between first and last column of Table 2).

H1N1 Concern Level	Not Received H1N1 Vacc Count	Received H1N1 Vacc Count	Grand Total	P
0	37,711,658	5,798,467	43,510,125	0.15
1	89,159,299	18,018,480	107,177,779	0.20
2	109,596,584	33,096,948	142,693,532	0.30
3	43,569,991	18,391,081	61,961,072	0.42
Grand Total	**280,037,532**	**75,304,976**	**355,342,508**	

Table 2- P Values for each H1N1 concern level for H1N1 Vaccine

For features with binary answers, changing in P value when the response changes from 0 to 1, is calculated as a measure to quantify the effect of that feature on one's receiving the vaccine. Table 3 shows this measure for all binary values for H1N1 vaccine and is ordered by features with greatest impact on receiving the H1N1 vaccine.

Binary Features	P[1]/P[0]
Doctor Recc H1N1	7.11
Health Worker	3.1
Doctor Recc Seasonal	2.79
Health Insurance	2.78
Behavioral Face Mask	1.74
Behavioral Wash Hands	1.74
Chronic Med Condition	1.64
Children Under 6 months	1.61
Behavioral Antiviral Meds	1.51
Behavioral Touch Face	1.5
Behavioral Avoidance	1.31
P[Married]/P[Not Married]	1.28
P[Own]/P[Rent]	1.25
P[Female]/P[Male]	1.12
Behavioral Outside Home	1.11
Behavioral Large Gathering	1.09

Table 3- Binary features impact on receiving the H1N1 vaccine

The fluctuation in receiving the H1N1 vaccine for HHS geo-region feature, which contains 10 geo-regions, as is shown in **Error! Reference source not found.**, are not statistically significant.

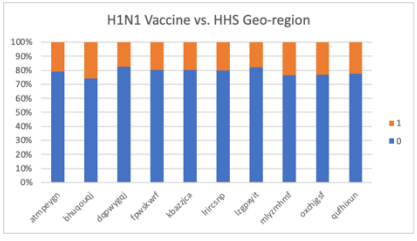

Figure 6- H1N1 vaccine vs. geo-regions graph

Thus, following features will increase one's likelihood of getting an H1N1 vaccine:
(Features with correlation values higher than or equal to 0.95 from Table 1):

- Higher levels of H1N1 Concern
- Associating higher risk with seasonal flu
- Higher levels of H1N1 knowledge
- Being above poverty line

(Features with P[1]/P[0] values higher than two from Table 3):

- A doctor recommends taking the H1N1 flu vaccine
- Being a health worker
- A doctor recommends taking the Seasonal flu vaccine
- Having health insurance

Seasonal Flu Vaccine

Similar analysis has been conducted for seasonal flu vaccine. See Table 4 for correlations and line coefficients of multiple-choice features with P variable.

Feature	Line Coefficient	Correlation
H1N1 Concern	0.28	0.996
Employment Status	0.42	0.99
Opinion H1N1 Risk	0.32	0.98
Opinion Seasonal Risk	0.68	0.98
H1N1 Knowledge	0.29	0.97
Income Poverty	0.21	0.94
Opinion H1N1 Vacc Effectiveness	0.3	0.93
Age Group	0.38	0.91
Opinion Seasonal Vacc Effectiveness	0.49	0.82
Education	0.07	0.58
Opinion H1N1 Sick from Vacc	0.02	0.29
Census MSA	-0.01	-0.2
Opinion Seasonal Sick from Vacc	-0.07	-0.4
Household Children	-0.15	-0.88
Household Adults	-0.16	-0.98

Table 4-Multiple choices features correlation to P for seasonal vaccine

For binary features a similar approach as Table 3 calculation is applied as well. See Table 5 for impact of binary features on receiving seasonal vaccine.

Binary Features	P[1]/P[0]
Doctor Recc Seasonal	5.33
Health Insurance	4.08
Doctor Recc H1N1	2.67
Health Worker	2.29
Chronic Med Condition	2.15
Behavioral Wash Hands	1.81
P[Own]/P[Rent]	1.74
Behavioral Touch Face	1.67
Behavioral Face Mask	1.41
Behavioral Avoidance	1.39
P[Female]/P[Male]	1.37
Behavioral Large Gathering	1.34
Behavioral Outside Home	1.29
P[Married]/P[Not Married]	1.22
Behavioral Antiviral Meds	1.07
Children Under 6 months	1.04

Table 5-Binary features impact on receiving the seasonal vaccine

HHS geo-regions does not significantly influence whether the respondents received the seasonal vaccine or not.
Thus, following features will increase one's likelihood of getting a seasonal flu vaccine:
(Features with correlation values higher than or equal to 0.95 from Table 4):

- Higher levels of H1N1 Concern
- Being employed
- Associating higher risk with H1N1 flu
- Associating higher risk with Seasonal flu
- Higher levels of H1N1 knowledge

(Features with P[1]/P[0] values higher than two from Table 5):

- A doctor recommends taking the seasonal flu vaccine
- Having health insurance
- A doctor recommends taking the H1N1 flu vaccine
- Being a health worker
- Having chronic medical condition

Notes

When creating predictive model, we should keep in mind that some of these features are also correlated. For example, people who are more concerned about H1N1 are generally associating higher levels of risk to this disease as shown in Figure 7. another example is that people who associates higher risk to H1N1 flu generally tend to assume higher risks for seasonal flu as well. This creates an interdependency between the features which makes it difficult to study each

feature's influence independently on the target variable. These types of interdependencies are also evident between employment status, health insurance and income poverty features.

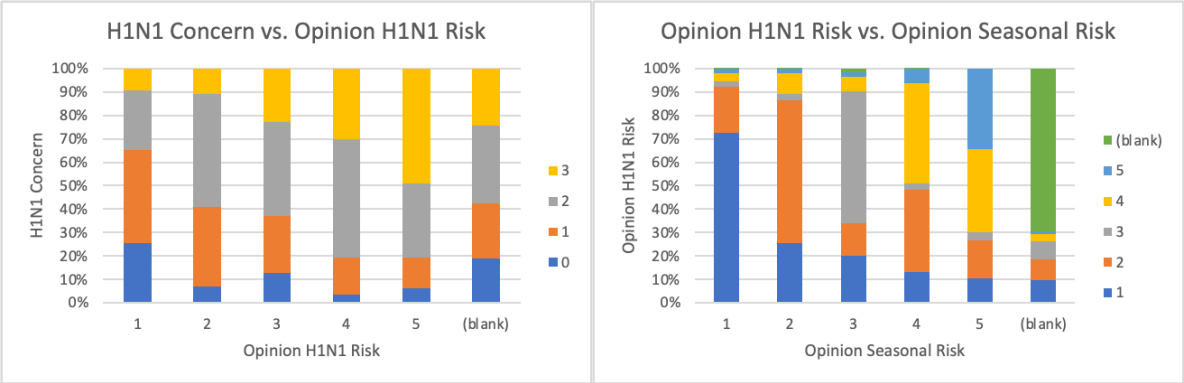

Figure 7- Graphs indicating interdependencies of some of the feature

One more point to note is that when a doctor has recommended each of these vaccines the likelihood of getting the vaccine has increased significantly. Despite this fact only 20% and 30% of the participants claimed that a doctor has recommended them taking H1N1 vaccine and Seasonal vaccine respectively.

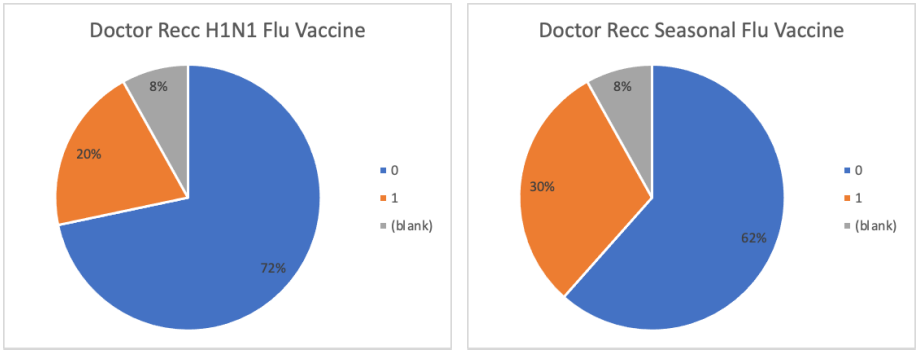

Figure 8- Vaccine recommendation by doctors

Conclusion

Based on the methodology section we can see individual opinion about the risk of the disease and the level they are concern about it plays a significant role in whether they decide to get vaccinated or not. That shows the importance of informing the public about the diseases and making sure they are aware of the risks associated with getting sick. This analysis also shows that doctors can be very helpful to get more people vaccinated just by recommending them to do so. Another factor that is crucial is having a health insurance and a good economic position. When dealing with contagious viruses like COVID-19, distributing the vaccine free of charge and without a need for health insurance will be helpful to get more people vaccinated. Related government agencies should also specifically target low-income families to ensure they are provided with options to get vaccinated.