Second Capstone Report
Behaviors Deterministic of Vaccine Acceptance

Problem Statement

While the most severe effects of COVID-19 finally appear to be waning, vaccine adoption is one of the most critical factors in speeding up this process and preventing additional virus variants from spreading. In 2009 an outbreak of the H1N1 virus or, “Swine Flu,” was the most similar crisis compared to what we are dealing with today.

Shortly after a vaccine was provided, the U.S. performed a phone survey and asked respondents whether they received the H1N1 or seasonal flu vaccine. In addition, questions related to their health, household, socioeconomic background, opinions and behaviors were asked.

Using this survey information, how well can we predict an individual’s likelihood of getting vaccinated? What behavior, opinion or demographic data can be focused on to help improve vaccine rates? If we can understand what drives individuals to get vaccinated there is a good chance we can apply the same insights to the COVID-19 pandemic.

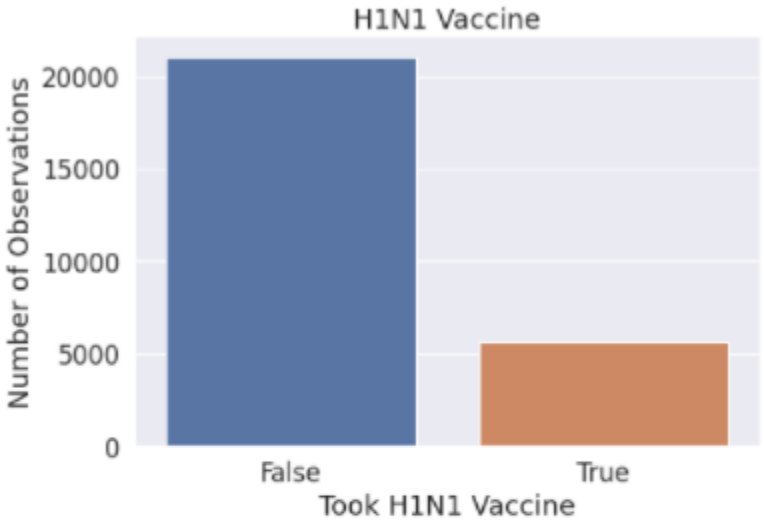
Data Wrangling

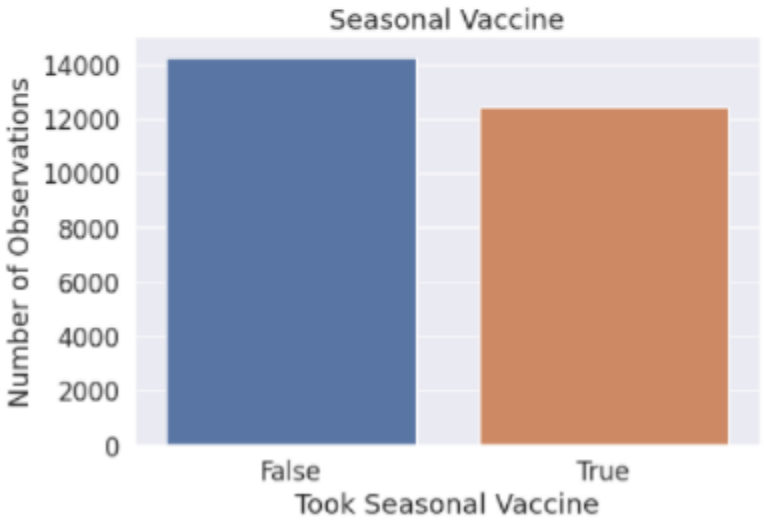
The dataset was taken from the Driven Data Flu shot prediction competition, which was taken from the CDC. I was fortunate in that most of the data was pretty clean to begin with. All features were categorical with only a few possible choices. There were three features, occupation, industry and geographic region that were encoded with arbitrary numbers. The actual industry or location that was represented by these numbers was never given. As well, two of these columns contained over 50% null values. Since this data is essentially useless, besides seeing that there was correlation with this occupation or industry we have no way of finding out, I dropped the columns.

Most of the invalid observations were in the columns I just dropped, for the remaining null values I treated them as an additional category. Considering the features provided had at most 6 categorical choices per feature, and that not making a choice is a choice in itself, this seemed like an acceptable course of action and allowed me to essentially preserve all the data besides the dropped columns. Since the choices for each feature were reasonable I was able to apply dummy encoding to the entire dataset. The only remaining cleaning tasks were to set all values to integer type and clean up some column names so that XGBoost wouldn’t have any issues when modeling. The data now being entirely binary there was no need for standardization. The final dataset I used to model was 128 columns and 26707 entries.

Exploratory Data Analysis

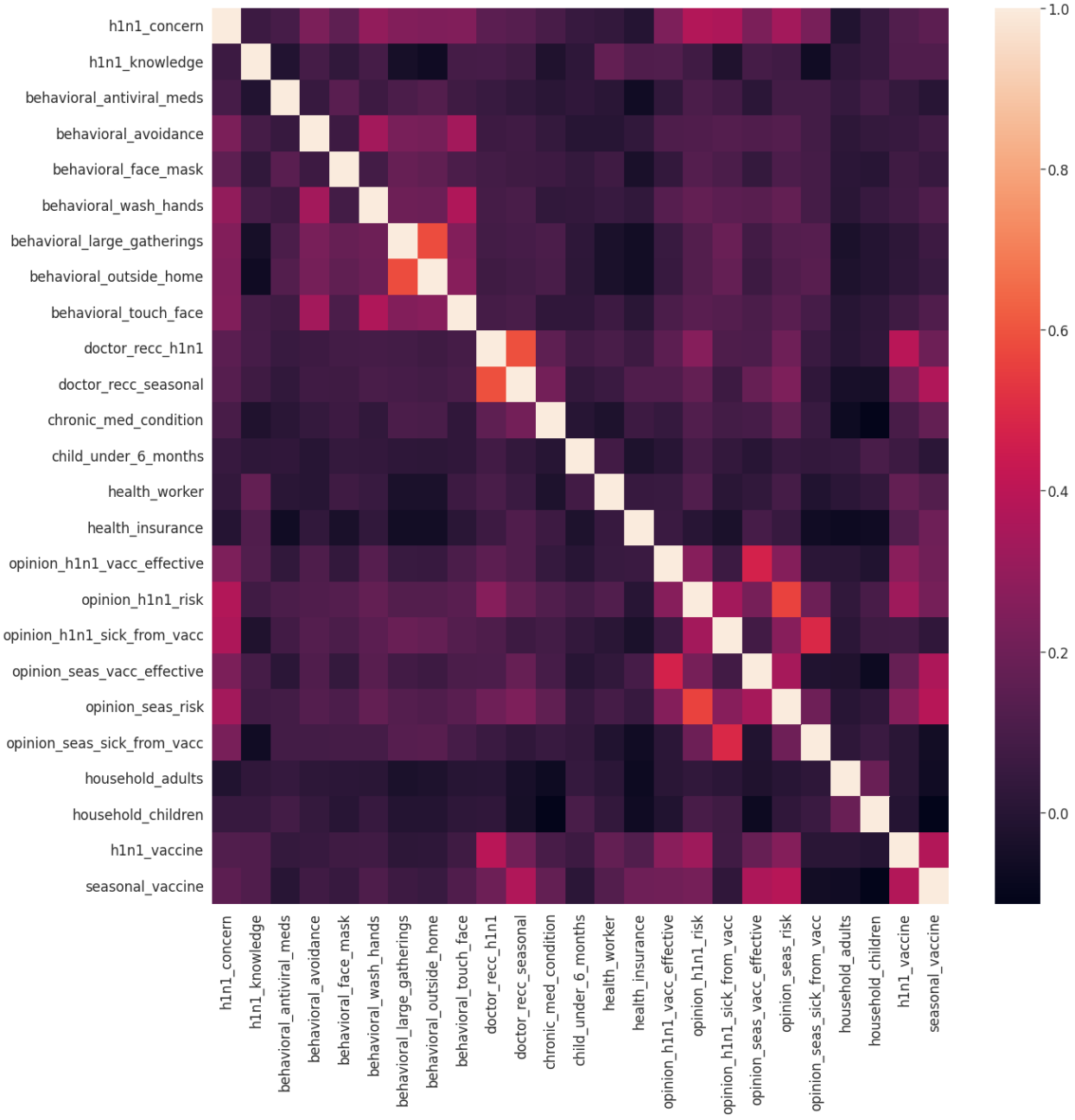
One of the first things I noticed was that although the results for the seasonal vaccine were pretty evenly split, results for the H1N1 were imbalanced almost 4:1 in favor of those not taking the H1N1 vaccine.

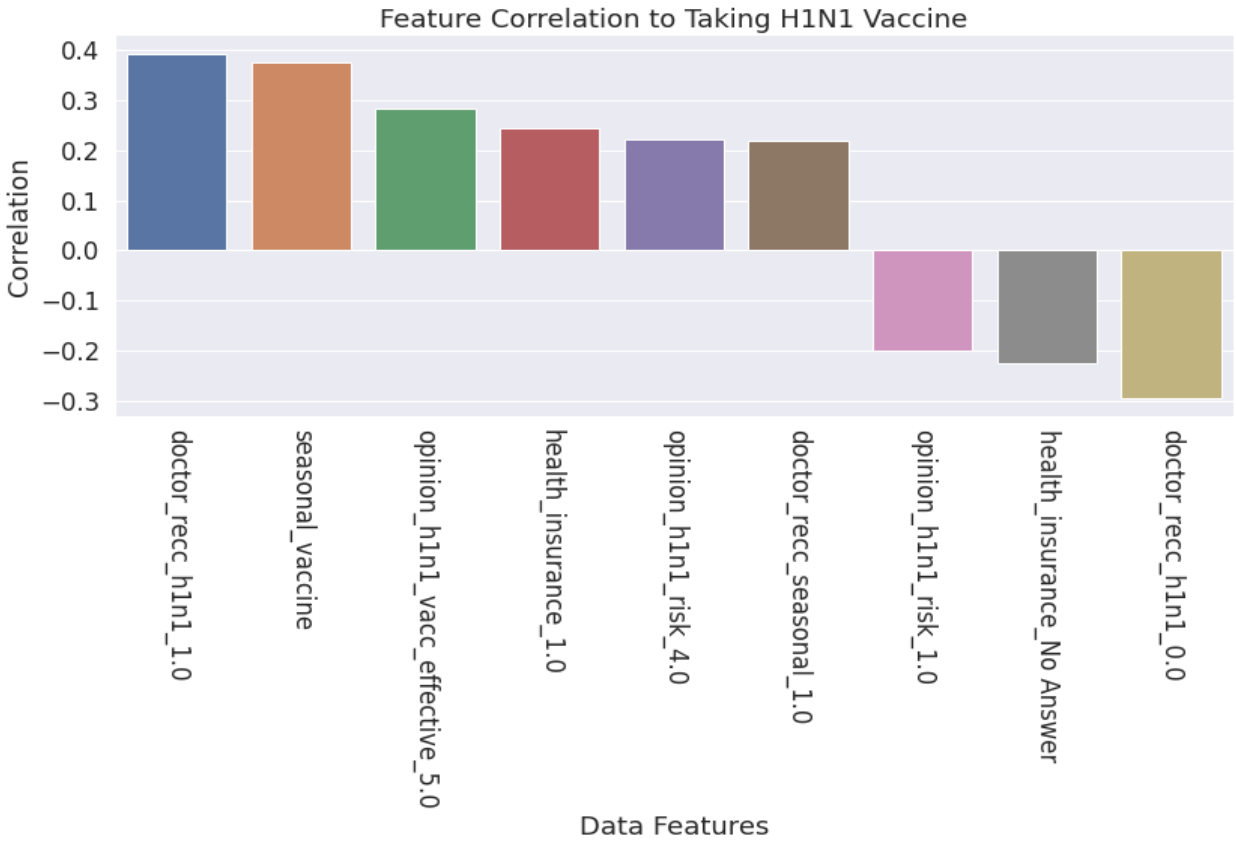


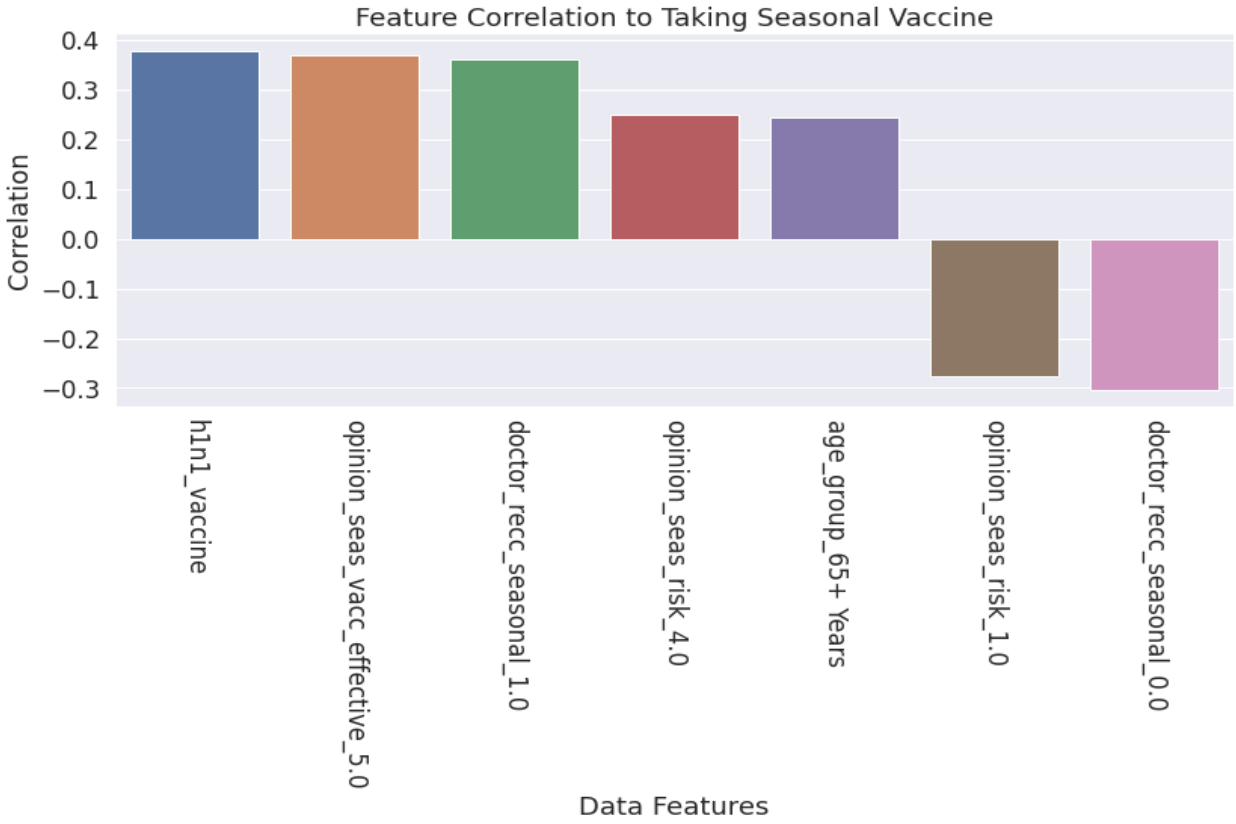


To manage the imbalance I'll choose modeling methods that are better suited to imbalanced data, as well as creating some synthetic data.

Since we are looking for what features best predict vaccine acceptance I focused on what features were most highly correlated to the targets. I generated the correlations and heatmap of the entire dataset before doing dummy encoding and stored features with either a correlation greater than 20% or less than -20%. Behavior is a type of feature where a negative correlation can be just as useful as a positive correlation if you're able to interpret the data properly.


h1n1_concern
h1n1_knowledge
behavioral_antiviral_meds
behavioral_avoidance
behavioral_face_mask
behavioral_wash_hands
behavioral_large_gatherings
behavioral_outside_home
behavioral_touch_face
doctor_recc_h1n1
doctor_recc_seasonal
chronic_med_condition
child_under_6_months
health_worker
health_insurance
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
h1n1_vaccine
seasonal_vaccine
1.0
0.8
0.6
0.4
0.2
0.0


Feature Correlation to Taking H1N1 Vaccine
Correlation
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
doctor_recc_h1n1_1.0
seasonal_vaccine
opinion_h1n1_vacc_effective_5.0
health_insurance_1.0
opinion_h1n1_risk_4.0
doctor_recc_seasonal_1.0
opinion_h1n1_risk_1.0
health_insurance_No Answer
doctor_recc_h1n1_0.0
Data Features

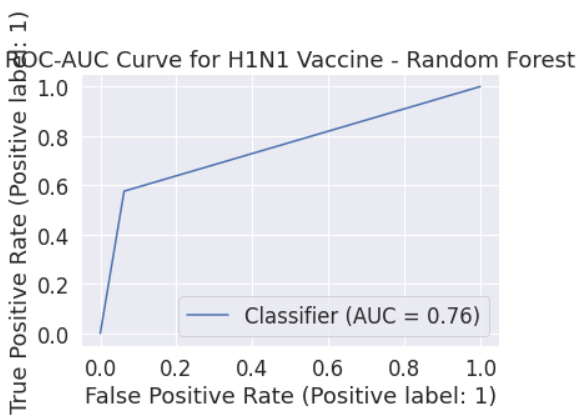

Feature Correlation to Taking Seasonal Vaccine
Correlation
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
h1n1_vaccine
opinion_seas_vacc_effective_5.0
doctor_recc_seasonal_1.0
opinion_seas_risk_4.0
age_group_65+ Years
opinion_seas_risk_1.0
doctor_recc_seasonal_0.0
Data Features

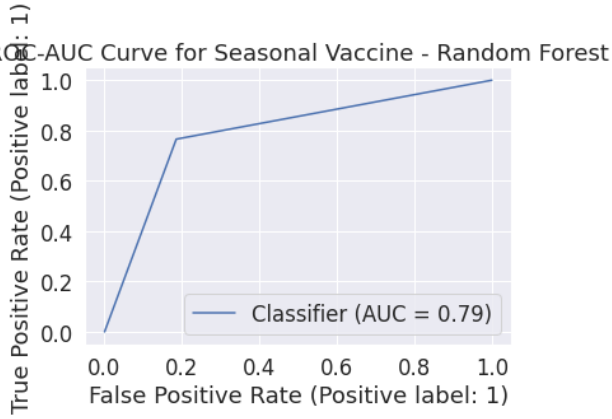
Modeling

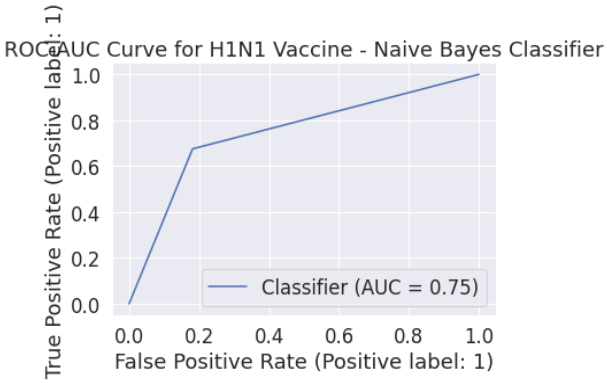
I needed to choose classification algorithms that were better at handling unbalanced data so I decided to go with Random Forest, Naive Bayes & XGBoost. I had a good feeling about Random Forest and XGBoost, being ensemble algorithms. I chose Naive Bayes because of its speed and XGBoost because of its reputation.

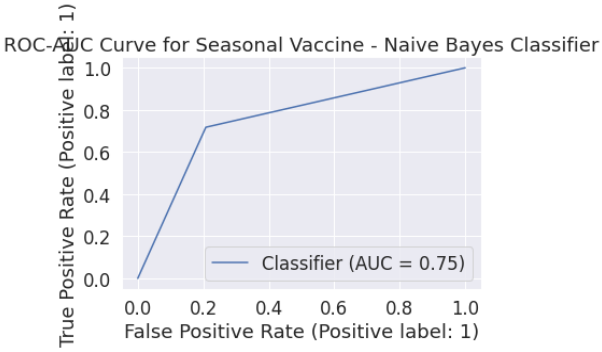
I performed a grid search for each algorithm and each target variable, h1n1 vaccine and seasonal vaccine, using the data before it was balanced synthetically. The synthetic balancing was only performed on the dataset for H1N1 results. I passed the best parameters into the models and produced predictions on the test split's.

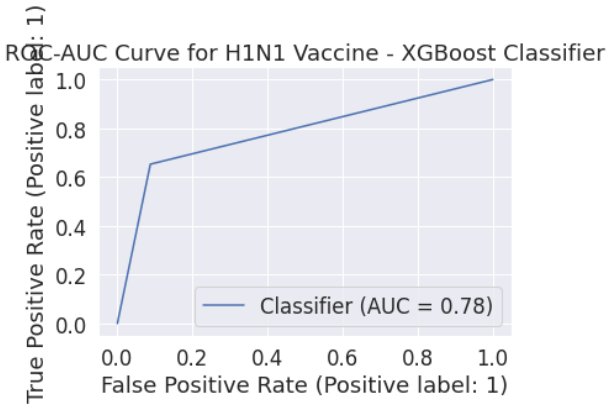
We don't particularly care about greater accuracy for True Positives over True Negatives. Since we are going for over-all best fit I used the roc-auc score to grade the different models. They were all very close but XGBoost was the winner. All 3 models had around a 0.75 AUC score, but XGBoost was able to achieve 0.78 for the H1N1 Label. Parameter tuning improved the AUC score for Random Forest and Naive Bayes by 0.01 but for XGBoost the defaults produced the same results. The parameters I had GridSearch parse were related to overfitting, however XGBoost's defaults worked quite well.

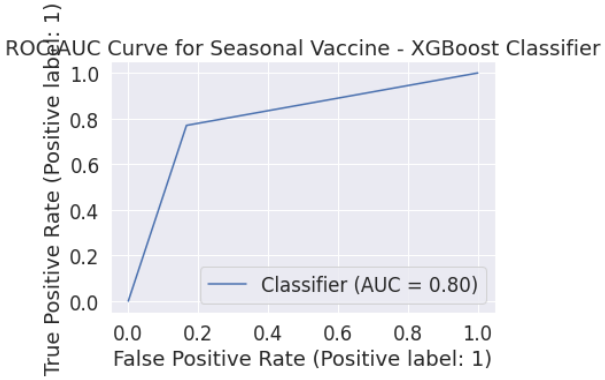




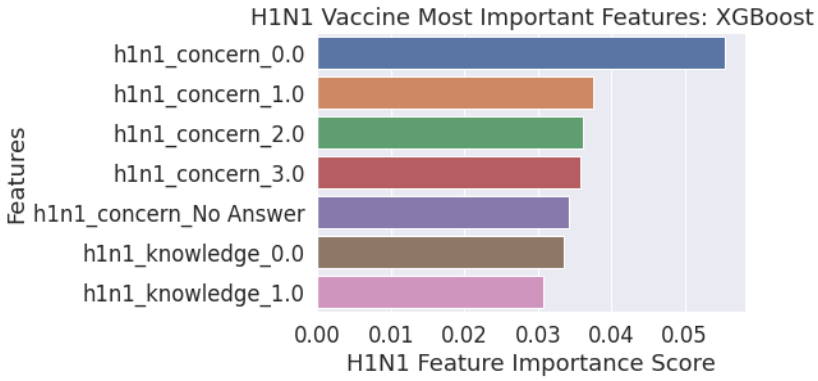


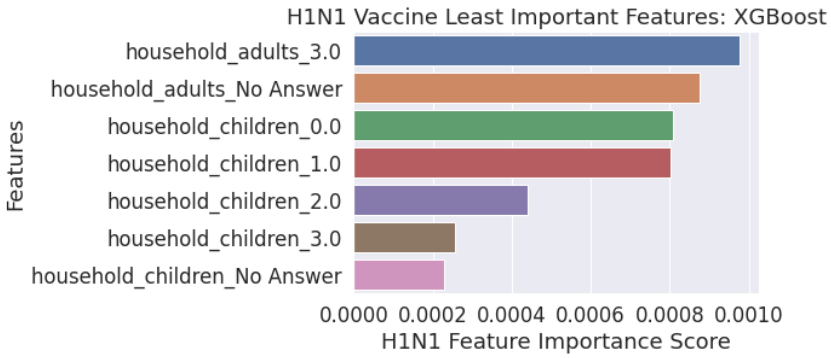






From the XGBoost classifier I extracted the 7 most and 7 least important features.





Oddly, the most and least important features were the same for both the H1N1 vaccine and Seasonal Vaccine so there is no need to show the features for the seasonal vaccine.

Analysis

The features from the pearson correlation and XGBoost's feature importance method appear very different but upon closer analysis validity can be logically inferred.

The most important factor in getting people to take the H1N1 vaccine appears to be a recommendation from one's Doctor. Next is whether or not they took the seasonal vaccine and after that, having the highest opinion on whether the vaccine is effective. It looks like the most important factor so far is trust. It makes sense that those who trust the vaccine more from either being used to getting the seasonal vaccine, professional deference and ultimately personal feelings. The next 3 features, having access to health insurance, feeling the H1N1 virus is a credible threat and one's Doctor recommending the seasonal flu vaccine, corroborate the first 3 features but add an additional element of fear. Access to insurance isn't surprising as cost of treatment is always a factor in deciding to see a Doctor, but it also implies that those with insurance more than likely see a Doctor more regularly and are therefore more trustful of getting vaccinated.

I also graphed the top 3 negative correlations, unfortunately they don't provide much additional insight but they strengthen the positive correlations as they are essential behavioral inverses of the features with the greatest positive correlation. Having a low opinion of the dangers of H1N1, declaring "No Answer" regarding insurance, which we can probably assume means no insurance, and one's doctor not specifically recommending the vaccine, are the greatest contributing factors to one not getting vaccinated.

The seasonal vaccine features were almost the same in implication. Trust of the vaccine, recommendation by their doctor and a high opinion of the danger of the seasonal flu were the primary drivers. There was one additional feature with high correlation, the individual being over 65 years of age. This could be due to seasonal vaccines being cheaper or more readily available for that demographic and their comfort in taking the seasonal vaccine. The H1N1 vaccine didn't share the same popularity with the same age demographic more than likely due to being relatively new. Understandably, citizens over 65 are less trustful of new things despite being in greater need of vaccination due to having less robust immune systems than the younger demographics.

Looking at the most important features generated by the XGBoost model, it seems to contradict our correlation results. The top 4 features measure one's opinion of the danger H1N1 presents, the highest correlation being "Not at all concerned" and decrease as concern increases. The Next 3 correspond to one having little to no knowledge about H1N1. These results seem contradictory to the correlation but taking the time to understand what best features mean to the algorithm, they make sense. A features importance is scored based on essentially how large a portion of the decision tree it's branch makes up for a specific result. In this case it implies that concern regarding dangers of the H1N1 virus are the greatest drivers for whether someone will get vaccinated or not. Thinking critically, this information supports the features from our correlation results.

Summary/Reccommendations

.

. Based on the data, the greatest drivers of vaccination rates are trust, knowledge and fear. This is based on a model that has close to 80% accuracy. Although no model can perfectly predict the actions of an individual, it illuminates trends and confirms attitudes we can see today that are affecting vaccination rates in the midst of the COVID-19 pandemic. The most important things that can be done to increase vaccination rates are to educate the public, engage GP doctors to educate and encourage their patients and strike a balance between fostering trust and accurately addressing the gravity of the situation. The last is a particularly slippery slope of which we're seeing the results of today.

Ultimately it all comes down to individual choice and no matter how hard you try some people will still refuse. However it appears that being honest with the public, educating the public truthfully and engaging family doctors, who have a face and relationship with their patients, is the best way to generate trust and increase vaccination rates