BUILDING MODELS TO PREDICT VACCINATIONS

By: Nora Trapp*
GB 656

*Analysis and code were done as a team by Rebecca Cox and Nora Trapp

Business Frame & Problem Statement:

I am the CEO for Madison County Health Department and my goal is to make the public's health a number one priority through strategic initiatives. With the increase of COVID – 19 and other illnesses, vaccines have been becoming more important than ever. However, vaccines are not easy to acquire and expensive to do so. Oftentimes, my team and I are unsure of the demand in our area, so we must guess on how many to order. This leads to us either throwing our excess capacity away, or turning down individuals because we ran out.

One way that we can fix this is through a strategy that allows us to know how many vaccines to order and how to allocate them more efficiently throughout the county. We have recently become aware of a machine learning model that can take a variety of information on demographic and behavioral features and predict the probability our citizens will get the H1N1 and/or flu vaccine.

By using this predictive model, we can see the specific areas that are less likely to get the vaccines and areas where the vaccines are in high demand. This will not only tell us where we need to invest more money in vaccine campaigns more efficiently, but it will also allow us to allocate the correct number of vaccines to each area. This will ensure cost savings by preventing waste, but also provide comfort that we will not run out and make sure every individual who wants to protect themselves has the opportunity to do so.

By having access to this new machine learning model, Madison County Health Department is better equipped to increase overall vaccination rates and ensure the health and well-being of the county is at its highest potential.

About the Data

The data used to build this model was from the National 2009 H1N1 Flu Survey conducted by the US in late 2009 and early 2010. This phone survey asked respondents whether they had received the H1N1 and seasonal flu vaccines, along with questions about themselves. These additional questions covered their social, economic, and demographic background, opinions on risks of illness and vaccine effectiveness, and behaviors towards mitigating transmission.

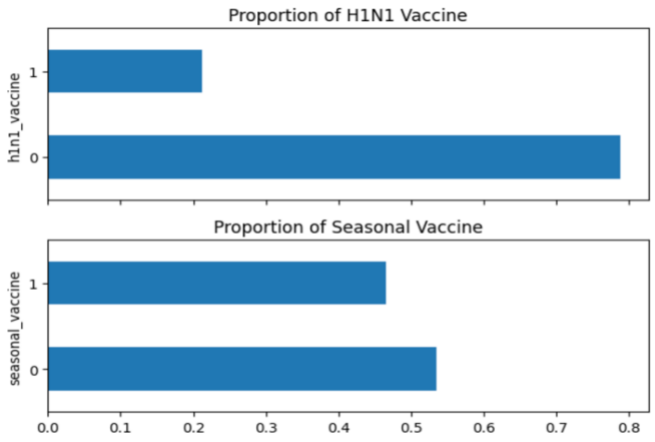
Approach:

Data Exploration:

Before building models on the flu shot data, data exploration must be done first to fully understand what the dataset says. First, we looked at how many rows and columns were in the dataset. In doing this, we found 26,707 rows and 35 columns. Each row corresponds to one individual and there are 35 features that predict two binary outcomes: if the individual got the H1N1 and if the individual got the seasonal flu vaccine. So, we are working with a classification problem. Next, we look at the datatypes of each of the features. We find that most are "float64" or integers, however there are 12 features that contain text in their fields.

Next, it's time to dig deeper into looking at relationships. First, we looked at the proportion of individuals in the dataset who got both vaccines. Through this graph below, we see that about 80 percent of

individuals didn't get the H1N1 vaccine and only 20 percent of the dataset did. With the seasonal vaccine, it is much more common and about 50 percent get it and 50 percent don't.

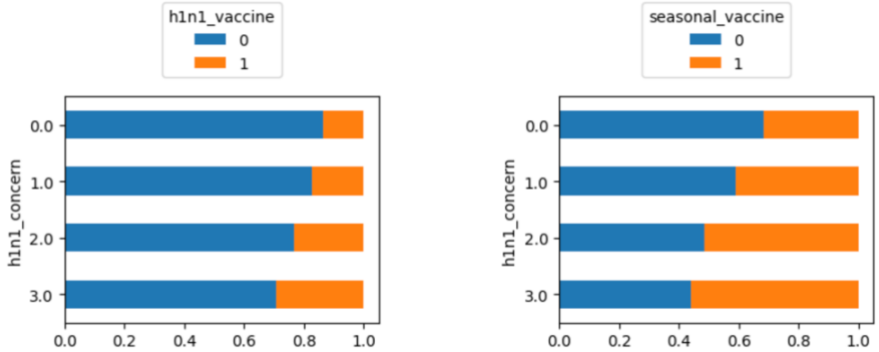


If we look at this next chart below, we see the percentages of people who got both vaccines, just one, or none.

seasonal_vaccine	0	1	All
h1n1_vaccine			
0	0.497810	0.289737	0.787546
1	0.036582	0.175871	0.212454
All	0.534392	0.465608	1.000000

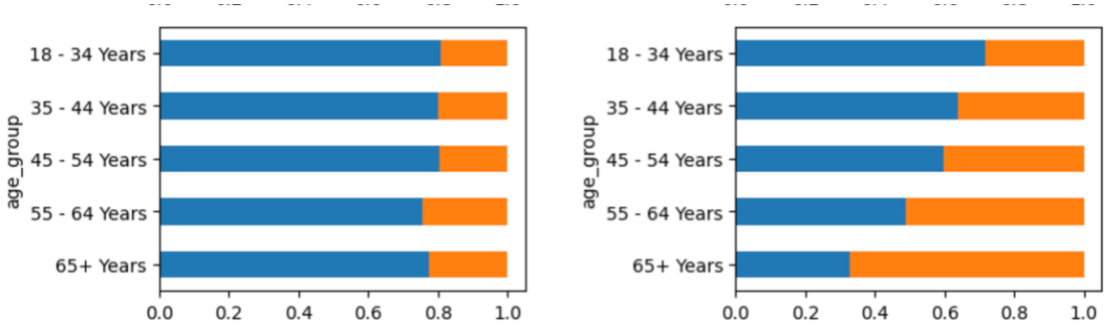
Here, we see that about 50 percent of the entire dataset didn't receive any vaccine. Only about 18 percent received both the H1N1 and the seasonal vaccine. About 29 percent of the dataset received the seasonal vaccine but not the H1N1, and only about 3 percent received the H1N1 and not the seasonal. With this information, we can conclude that only about half of the dataset is being vaccinated and that if people are going to be vaccinated, it is more likely it'll be the seasonal vaccine and not the H1N1.

The last step in our data exploration phase is looking at the relationship between a variety of features and our outcomes. First, we started with the feature "H1N1_concern."



The graph on the left shows us that as the H1N1 concern increases, more people are getting the H1N1 vaccine. On the right, it also shows that as the H1N1 concern increases, more people are also getting the seasonal vaccine. This is most likely because if their H1N1 concern increases, they're more likely to get the H1N1 and if they get that one, they might as well get the seasonal vaccine at the same time.

The graphs below show the proportion of individuals getting each vaccine by age group. With the H1N1 vaccine, there's isn't much difference between age group. However, with the seasonal vaccine, it looks like as people age, they're more likely to get the seasonal vaccine.



Data Cleaning:

After we explored the data, we realized we needed to take some additional steps to clean it before building models. First, we dropped three features that had more than 20 percent of the values missing (health insurance, employment industry, and employment occupation). Then we converted our 12 features that were categorical to dummy variables. During this step we also scaled our data since some of our models require that. Then, we used the k nearest neighbors technique (with n neighbors equal to 5) to impute the remaining N/A values in our dataset. This method works by filling in missing values by preforming local regression on the 5 most similar individuals in the dataset. Last, we split our training data into a training and validation set.

Constructing and Evaluating Models:

The performance of each model will be evaluated according to the area under the receiver operating characteristic curve (ROC AUC) for each of the two target variables.

Due to our dataset predicting two y variables rather than one, we build two models (one for y1 and one for y2) for each model below. We did attempt to try and build one model solely based on flu shots and ignored H1N1 for the time being. Then used this model with flu as an x variable as an "engineered feature" to predict H1N1 vaccination. We found that using this strategy didn't change our AUC much compared to just building two models. So, we decided to use the strategy we were most familiar with and build separate models for each of the y outcomes.

1. **Logistic Regression**:

   We chose to start with the logistic regression model since it's simple to build and a common starting place when building a variety of different models. The y1 (H1N1 vaccine) produced an area under the curve of .8246 and the y2 (seasonal vaccine) produced an area under the curve of .8534. No tuning is required for this model.

2. **LASSO Regression**:

Next, we ran a logistic LASSO regression because our dataset has many features and a large n. The tuning parameter in this model was C (like lambda in linear regression) and we ran a model for y1 and y2 iterating over a wide range of C's. For y1 the highest AUC was .825 and for y2 the highest was .853.

In addition to the LASSO regression, we also planned to run a ridge regression. We realized that in this case, our logistic regression was set to penalize the same as ridge regression, so the two produced the same AUC. From this finding, we concluded running a ridge regression wasn't necessary.

3. **Random Forest**:

Next, we ran a random forest because these models don't require much tuning and tend to predict very well. We still experimented around with the model and tuned a bit to see what would give us the highest AUC. For y1, we set our criterion to entropy (performed better than gini), with a max of 8 features, a total of 500 estimators, a min split of 5, and a min impurity decrease of .000001. This produced an AUC of .8311. For y2, the criterion was also entropy, the max features were sqrt (the default setting) which is essentially 7 because we have 50 parameters after scaling the data, the min split was again 5, estimators were also 500 and the min impurity decrease was .00000001. These parameters produced an AUC of .8562 for y2. To get these specific parameters, we tried many different combinations of many different parameters to find the ones that produced the highest AUC for each model So far, this model predicts the best for both y1 and y2**.**

4. **Gradient Boosted Tree:**

Next, we decided we wanted to try an ensemble model to see if that would increase our AUC even more. Again, we played around with the tuning to see what parameters would produce the best AUC. For y1 and y2, we set the n estimators to 5000 and the learning rate to .005. For y1 we set max depth equal to 4 and 3 for y2. We also added the tuning parameter sub sample (which is the fraction of samples used to fit the individual base learners) and set that to .1 both target outcomes. Y1 produced an AUC of .8345 and y2 produced an AUC of .8622. So, we see that this model did even better for both outcomes than the random forest.

5. **Neural Net**

Last, we decided to run a neural net to see one last time if we could improve our predictions. This model requires the most tuning, so again, it was trial and error to tune and see if we could improve the AUC. Both the number of layers and the neurons in each layer were decided by trying different numbers and running the model over and over. We also experimented with the activation functions in each layer and the learning rate. We set the dense layers and activation functions the same between each model. Both had a learning rate of .001, a batch size of 100, and 10 epochs. After 10 epochs, we started to overfit. Y1 produced an AUC of .8254 and y2 produced an AUC of .8546. So, we see that even though this model is the most complex of all, it still didn't predict the best.

Results:

Model Used	Y1 AUC Value	Y2 AUC Value
Logistic Regression	82.46%	85.34%
LASSO Regression	82.50%	85.30%
Random Forest	83.11%	85.62%
Gradient Boosted	83.45%	86.22%
Neural Net	82.54%	85.46%

The gradient boosted model preformed the best on predicting both y1 and y2.

The last step in the process is to predict on the test data using the gradient boosted model. Just like we did for the training data, there were necessary cleaning steps to take. First, we dropped the same three columns that had missing values. Then, we again converted the 12 categorical variables to dummy variables. Then, we used the k nearest neighbors' method to impute the remaining columns' missing values and replace them. Last, we predicted the probabilities for both target outcomes using the test data for everyone in the test set. Now, we officially have predicted the probabilities of individuals getting the H1N1 and flu vaccine on individuals outside our training data.

Conclusion:

I, as the CEO of Madison County Health Department, can now use the best predictive model, the gradient boosted model, to better understand the probability that an individual in the area will receive the H1N1 vaccine or seasonal flu vaccine. Having this information will help in two ways. First, it will allow me to see where the demand for vaccines is lower and gives me areas to target for vaccine campaigns to increase awareness. Second, it will ensure that I order enough vaccines and distribute them accordingly to prevent waste by throwing away any excess.

Significantly, the application of this model holds the potential for extension to other vaccines for the area. Considering the increasing prevalence of COVID-19 vaccines, leveraging this model presents an opportunity to anticipate the likelihood of individuals opting for that vaccination. Collection of data would need to take place that again measures a variety of behavioral and demographic features with a binary outcome variable of whether the individual got the COVID vaccine or not. Once this data collection phase is concluded, I could use this model in a similar way. Equipped with the appropriate dataset, this model can be adapted for many diverse vaccines, thereby enhancing the capabilities of Madison County Health Department in delivering optimal care to their citizens.