Springboard Data Science Capstone Project: Predicting Flu Vaccination Status

Thomas Stern

May 1st 2022

Contents:

1. Introduction:

In a time full of misinformation and conspiracy theories, imagine if we could predict which patients will get vaccinated based on a series of simple survey questions. If we could predict which patients get yearly seasonal flu vaccines, as well as vaccinations to novel viruses, we could target our efforts with much more precision. This could result in far more individuals becoming vaccinated making herd immunity a real possibility. As COVID-19 continues to wreak havoc on the world, it is becoming more and more important to target our vaccination efforts so that they result in the most possible vaccinations within the population. Ideally, with progressing medical technology and better aimed public health efforts, we can slow future pandemics and better deal with endemic viruses.

Now imagine, a novel virus spreading across the United States as well as the rest of the world killing hundreds of thousands of people in its first year. Sound familiar? Although less deadly than COVID-19, the H1N1 flu virus is a decent model for how people respond to a novel virus and its respective vaccine. In 2009 and 2010 the CDC conducted the National 2009 H1N1 Flu Survey, which consisted of various questions relating to demographics, opinions and behaviors surrounding seasonal influenza and the H1N1 “swine flu”. The survey also asked whether these respondents had received vaccines for each virus. Using this survey data I hope to create a model that can predict whether a person will get an annual flu vaccine using only the questionnaire provided. With the questionnaire answers and seasonal flu vaccination status, I will then be able to make an even more robust prediction for a novel virus/vaccination, which in this case is H1N1 (aka Swine Flu). I hope to be able to use these models for any new pandemic that arises so that future public health efforts can be made with optimal precision.

2. Data Wrangling and Cleaning:

All data for this project was obtained from DrivenData.org, who obtained it from the CDC’s National Center for Health Statistics (NCHS) and National Center for Immunization and Respiratory Diseases (NCIRD). Jointly, these groups sponsored The National 2009 H1N1 Flu Survey, which asked randomly-telephoned individuals a variety of questions along with their vaccination status between October 2009 and June 2010. The survey consisted of 35 questions: 13 yes/no questions about behaviors, health and demographics, 8 opinion questions rated 0-5 and 14 demographic questions with categorical answers. The training data that will be used to create the models consists of 26,707 respondents. The final model will then predict the probability for vaccination for 26,708 respondents in test data (does not contain true label of vaccination statuses).

The data was already relatively clean but there were a number of steps I performed in order to best visualize the data and also to be able to use it in machine learning models later on. First, any ordinal or binary variables were mapped to integers. This will help in visualizing the data as well as be sufficient for machine learning models. Next, each non-numeric column was mapped to an integer so that visualizations could be made quickly, to better understand distributions within the data. Another dataframe was made that kept each categorical variable as strings/objects so that later on we could one hot encode them for machine learning. For both dataframes, there were a number of missing values that also had to be dealt with. Most features were only missing a small proportion of data and filling them would be fairly straightforward. There were 3 columns that were missing close to half of their data, so these took a bit more thought when filling.

The health insurance feature asked if the respondents currently had insurance and the only possible answers were yes or no. Many respondents did not answer this question. It could be assumed that not answering this question meant they were more likely to be uninsured, but when looking at the national percentage of insured individuals, it seemed more likely that most of the missing data belonged in the "insured" group. Because there were many assumptions being made to come to either conclusion, I decided to make a 3rd category: "unknown". This would make this feature nominal categorical in the end, as there is no clear order to insured, uninsured and unknown.

The employment_industry and employment_occupation features also were missing a large proportion of data. They allowed for the respondents to identify with 1 of roughly 20 different categories (represented here as jumbled strings of letters to preserve anonymity). It was possible that those not answering this question did not feel that they belonged in any of the categories so for these features, I also created another category: "unknown".

For the rest of the data that was not missing large portions of responses I decided to use the mode of each column to fill missing information. I chose the mode, as opposed to the median, because most of the features did not show a normal distribution. Many, especially those involving opinions, were polarized to either high or low numbers as fewer people had neutral views/opinions. Mode seemed to better preserve the distributions seen in the raw data.

3. Exploratory Data Analysis:

Using the numeric data frame created during cleaning/wrangling, I was able to efficiently show all of the feature distributions as well as relationships between features and target variables. The following conclusions were drawn from our data:

3.1 Target Variables:

The two target variables in question were the yes/no (1/0) answers to the vaccination status questions for seasonal flu and H1N1 swine flu. **Figure 1** shows distributions of each of these variables. Seasonal flu vaccination rates were moderately higher at 46.6% of respondents vaccinated, whereas H1N1 was only 21.2%. This is to be expected as the seasonal influenza vaccine is well established as safe and effective whereas the H1N1 vaccine was brand new and only used for this one year. This will also make it better as a model for future pandemics as the vaccination rates are likely to lower for the novel virus at first, which leads to unbalanced data. We want our model to be able to handle an unbalanced data set when predicting whether an individual will get the new vaccine.

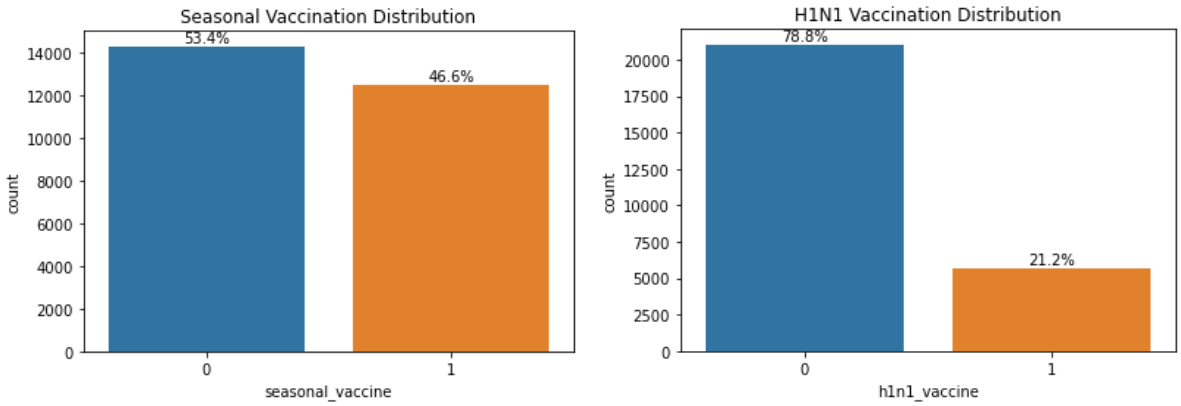


Figure 1: Distributions of seasonal and H1N1 flu vaccination status responses

3.2 Feature Variables:

After looking at our target variables, I next looked at how each feature correlates with the other features and with the target variables. To do this relatively quickly, I used heatmaps to visualize the correlations between the binary and ordinal data. Our data frame was too large to visualize all at once so I broke it down into 2 smaller sets, each containing the target seasonal and H1N1 flu vaccine variables. As seen in **Figure 2**, it seems there are no extremely strong correlations between any 2 independent variables and there are only minor correlations between independent variables and response variables.

Some variables to keep an eye on moving forward are the presence of Dr recommendations, beliefs about risks/effectiveness of vaccines and age group's effect on vaccination rates. **Figure 3** shows the largest correlations seen within the data, however the highest correlation coefficient was only 0.58, which is relatively low. This leads us to conclude that no single feature is going to determine the vaccination status of the individual. It will likely take the combination of all of these features for the

machine learning models to make robust predictions. Although somewhat disappointing in terms of exploratory analysis, this data is likely a great candidate for machine learning models that go far beyond the scope of an individual's ability to pick up on correlations in the data.

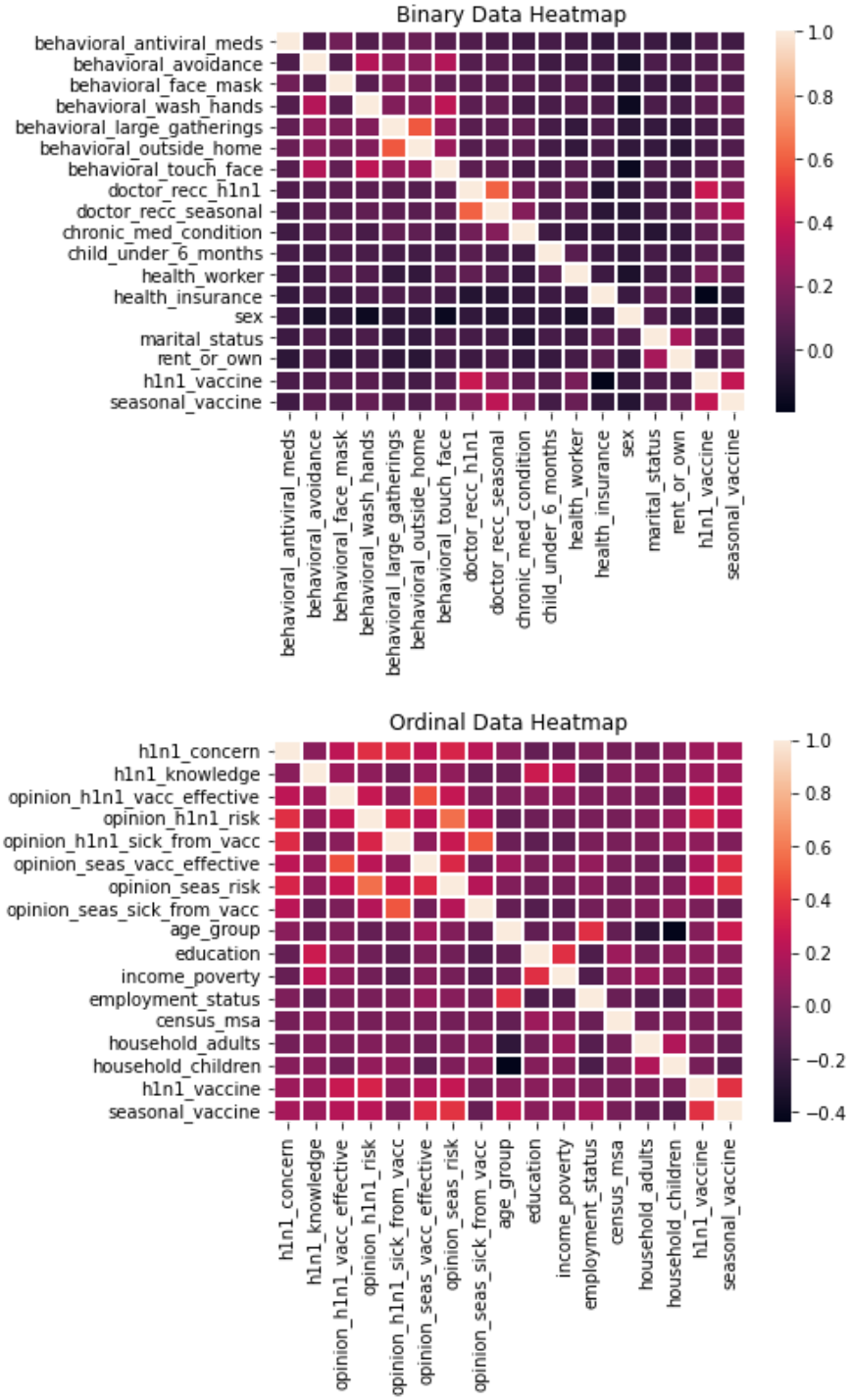
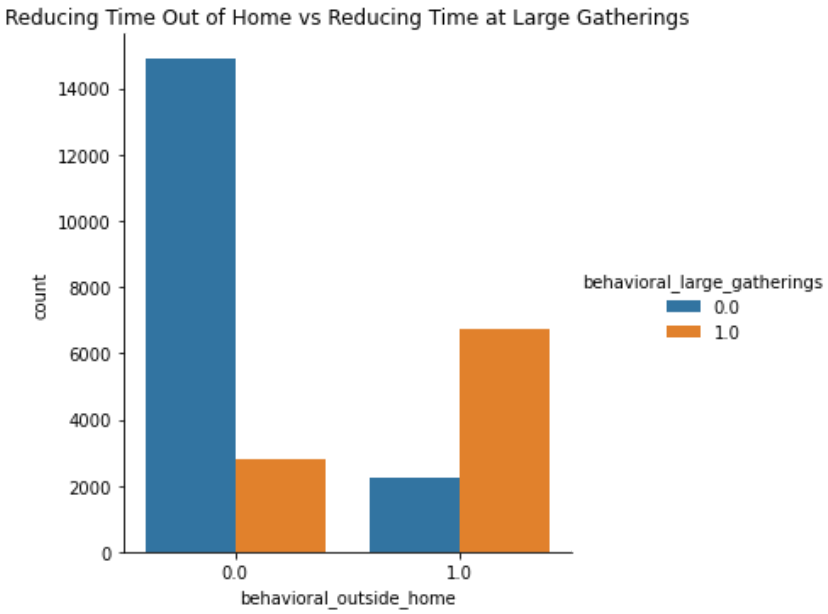
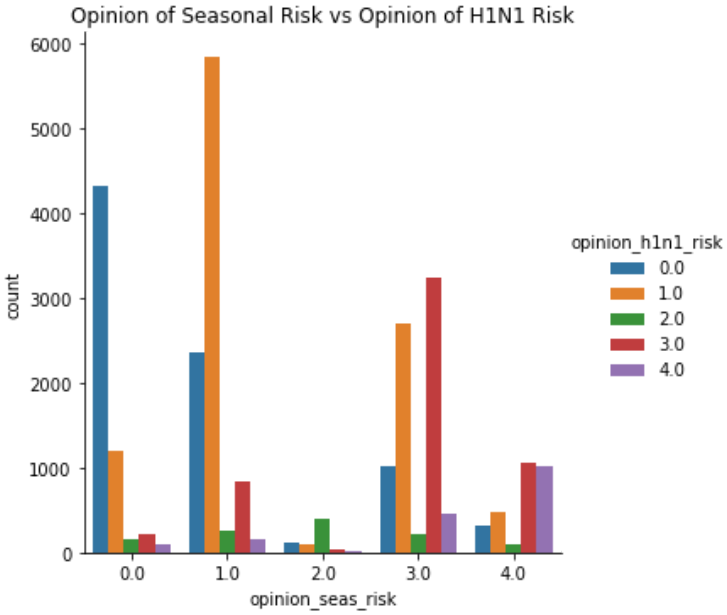
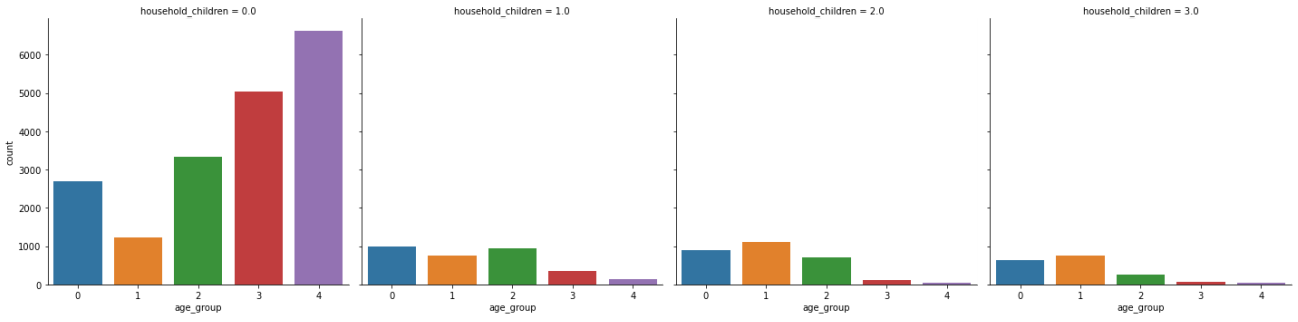


Figure 2: Heatmaps displaying correlations between features


Reducing Time Out of Home vs Reducing Time at Large Gatherings
count
behavioral_outside_home
behavioral_large_gatherings
0.0
1.0


Opinion of Seasonal Risk vs Opinion of H1N1 Risk
count
opinion_seas_risk
opinion_h1n1_risk
0.0
1.0
2.0
3.0
4.0


household_children = 0.0
household_children = 1.0
household_children = 2.0
household_children = 3.0
count
age_group

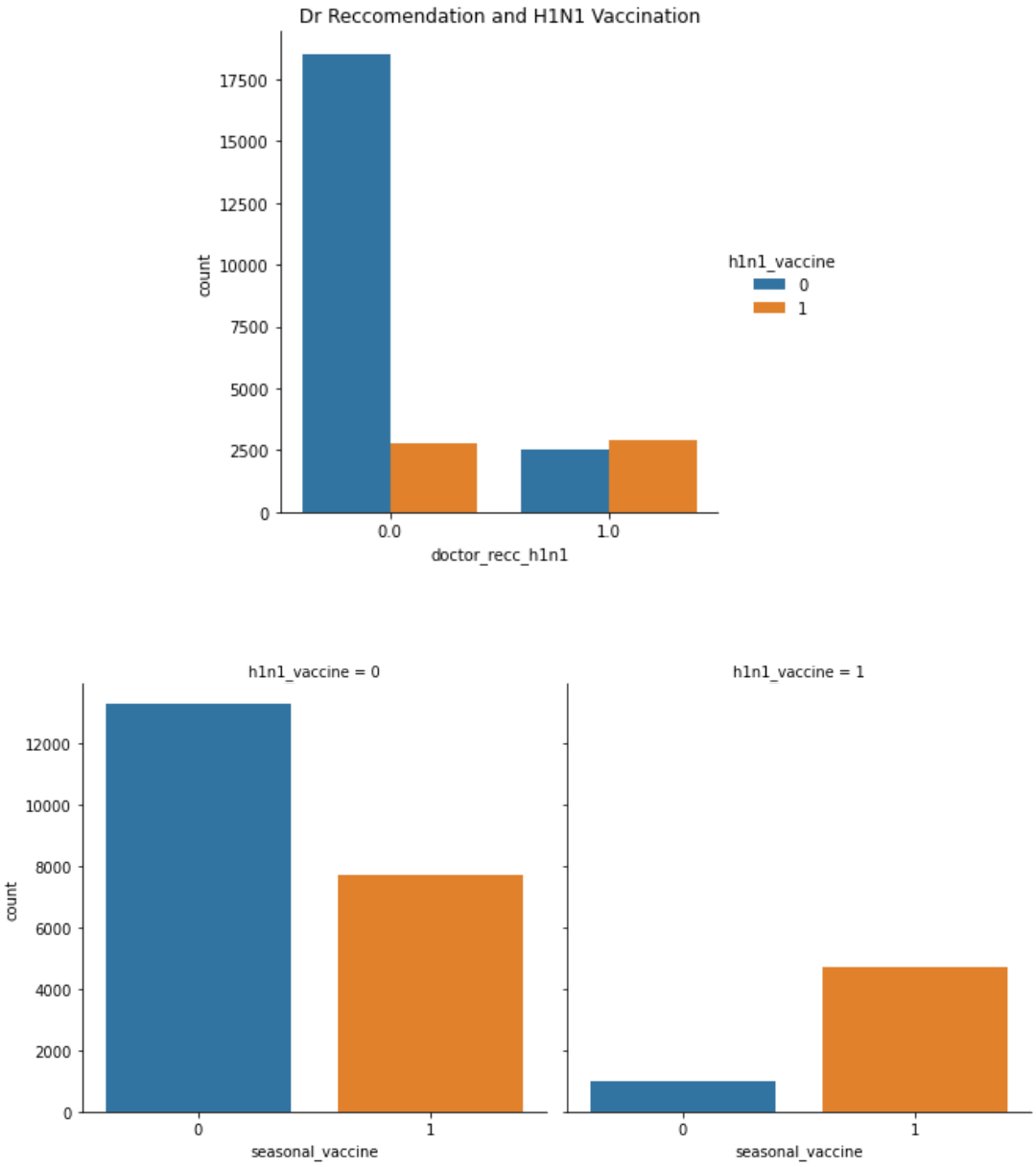


Figure 3: The strongest correlations are depicted above. In descending order, they had the following respective Pearson Correlation Coefficients: 0.58, 0.56, -0.44, 0.39, 0.37.

3.3 Analysis Conclusions:

Although we were not able to find many compelling correlations, some conclusions can be drawn about the demographics of the population being sampled. This sample is predominantly white, female, educated and above the poverty line. The large majority have no children in their household and limited contact with children under 6 months of age. This may not be a perfect sample of the entire population but hopefully it is close enough that our model will extend far beyond this sample.

4. Modeling:

From EDA, we were able to see that limited correlations between features and other features/targets may make this problem very well suited to machine learning models that can look beyond the obvious to make robust predictions. The final model must be able to use the feature data to reliably predict the probability that an individual is vaccinated against seasonal influenza and H1N1 swine flu. This only leaves a number of possible models as it must be a classification model that also has a predict_proba() method (gives probability), as opposed to just the predict() method which gives only the predicted class. The majority of these are decision tree based models but also include logistic regression, Naive Bayes and SVCs. I decided to compare logistic regression to a few different decision tree based models. The first 2 I compared were a random forest classifier and gradient boosting classifier.

Before inputting the data into various models it was necessary to preprocess it. This data was relatively balanced and did not require resampling via undersampling/oversampling. Scaling is not necessary for the decision tree classifiers but is for logistic regression. Although mostly on the same scale, the data was all scaled to be between 0 and 1 for the logistic regression model. All models, however, required one hot encoding for the nominal categorical variables. This was accomplished using pd.get_dummies(). All models also needed data to be split into train and test sets. I chose a test size of 0.3 and therefore a train size of 0.7. Also an additional column was created from the household_children and household_adults features. By adding them together, I created a household_total column which was seen to be more significant than either column alone later on.

Here I decided to use H1N1 vaccination prediction as something we can use as a model of novel vaccines as new viruses and pandemics arise (cheifly the COVID19 pandemic). I will therefore be able to use the seasonal_vaccine as an additional feature in predicting H1N1 vaccination status. In the future we will have this data as seasonal flu vaccinations are recorded in state and federal databases or can be self declared by those answering survey questions. Using H1N1 vaccination status to predict seasonal vaccination is less useful because the H1N1 vaccine was only given one year, whereas seasonal flu vaccines are given every year. Also it is more important to be able to use this information to predict who will get vaccinations for novel viruses during a pandemic.

4.1 Random Forest Classifier:

Using GridSearchCV I was able to determine which hyperparameters to use for a random forest model. For both the seasonal model and the H1N1 model, the only tuning necessary was to set criterion = entropy and min_samples_leaf = 4. The other hyperparameter that could be tuned was n_estimators, which defaults to 100 and gives reasonably accurate results. As you increase n_estimators beyond 100, there are marginal increases in accuracy but they come at the cost of rather long CPU times. In order to stay at the same magnitude of CPU times as logistic regression or gradient boosting, we must leave n_estimators near 500. My seasonal flu model showed a mean accuracy of 0.7829, while the H1N1 model achieved a mean accuracy of 0.8622. Using cross validation the random forest models were able to achieve mean ROC_AUC of 0.8569 and 0.8848 for seasonal and H1N1 respectively. The average of the 2 ROC_AUC was 0.8709, which will be the final metric for the contest.

I was also able to see which features were most important to the random forest models using the feature_importances_ attribute. They were relatively similar for both models. The top 5 features for the seasonal model were: opinion_seas_vacc_effective, opinion_seas_risk, doctor_recc_seasonal, age_group, opinion_h1n1_risk. The top 5 features for H1N1 were: seasonal_vaccine, doctor_recc_h1n1, opinion_h1n1_risk, opinion_h1n1_vacc_effective, opinion_seas_risk. This can be better visualized in **Figure 4.** Both models give a significant importance to many of the opinion survey questions (opinions on effectiveness, risks, and side effects). Interestingly opinions about the other vaccine seem to be important for both seasonal and H1N1. Doctor recommendations for each vaccine also show large importances for each vaccine respectively. Age group also showed significant importance for both models. Education and relative population density also made the top 15 for both models. Interestingly, there was a noticeable difference in importances in the health_insurance features. Health_insurance_1.0 and health_insurance_unknown both showed significant importance in the H1N1 model but barely made the top 20 for the seasonal model.

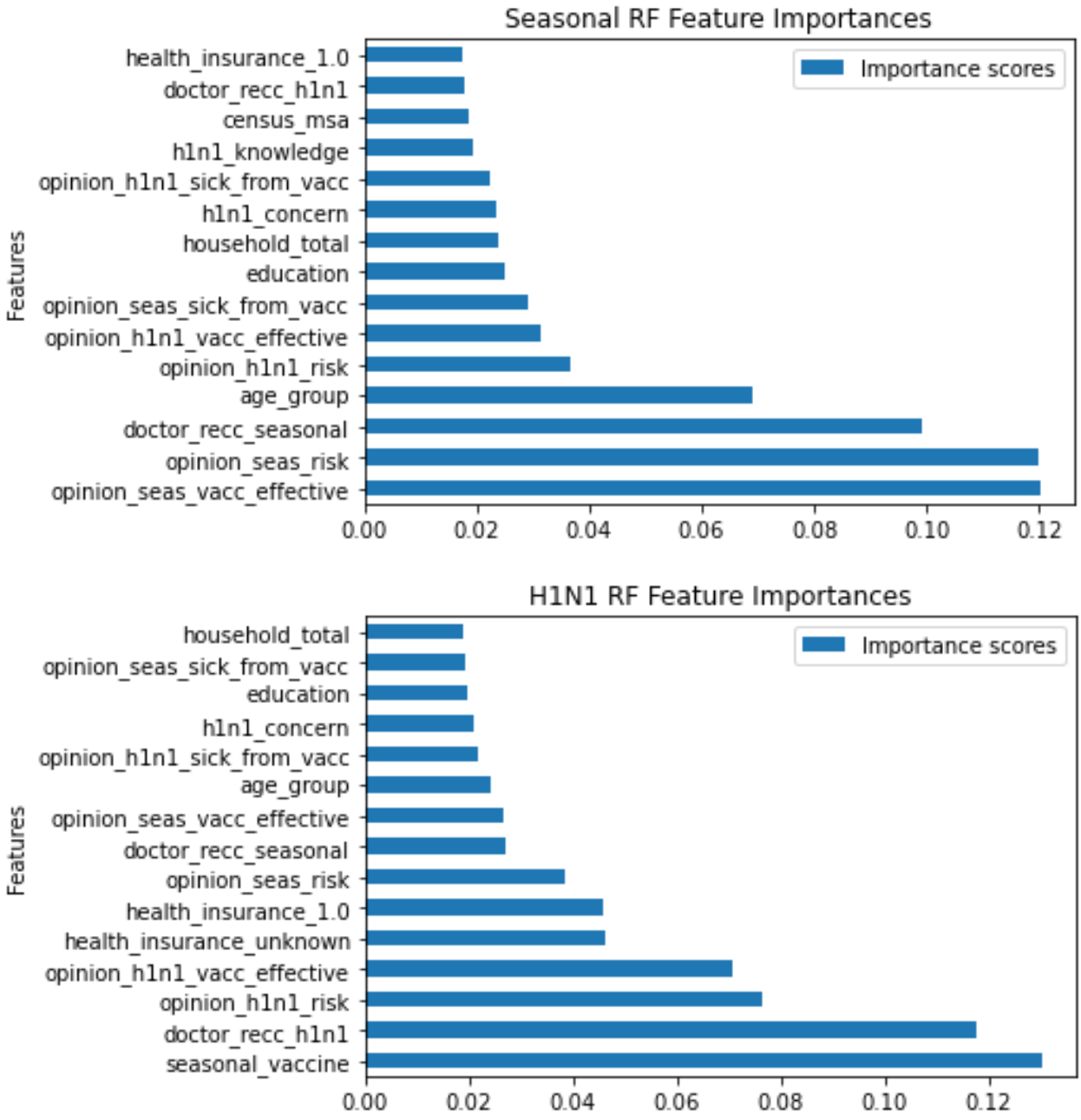


Figure 4: The 15 most important features in each Random Forest model. Most important features at the bottom of each graph

4.2 Logistic Regression:

Next, in order to see how this data would fit to a completely different model type, I wanted to confirm that I was headed in the right direction using a decision tree based model. For this model, cross validating hyperparamters was very important. Unlike RandomForest, GridSearchCV using LogisticRegression as the 'estimator' produced very different results for seasonal flu vaccination feature data and H1N1 vaccination feature data. With max_iter set to a relatively large number of 5000, I optimized 3 other hyperparameters. This allowed every model to fully converge before ending the GridSearch. I optimized 'penalty', 'C', and 'solver'. Penalty reduces the least contributive variable in the model towards a coefficient of 0. I tested None, l1, l2 and elasticnet. C refers to the strength of regularization with smaller numbers causing stronger regularization. The solver hyperparameter chooses which algorithm should be used to solve the optimization problem. Many of these parameter settings do not work in concert so many errors were thrown but of those that worked, I was able to get specific tuning for each model.

The seasonal vaccine model optimized with penalty = l1 and solver = saga. C was left at the default value of 1.0. With these hyperparameters, the model showed a cross validated accuracy score of 0.781 and a cross validated ROC_AUC score of 0.8554. The H1N1 vaccine model optimized with penalty = l1 and solver = 'liblinear'. C was also left at the default of 1.0. With these hyperparameters thew model displayed a cross validated accuracy of 0.8615 and a cross validated ROC_AUC of 0.8843. The average of the 2 ROC_AUC's was 0.8699, which was marginally lower than the random forest model. The random forest model could also increase its score further by increasing the n_estimators hyperparameter, whereas logistic regression was going to be at its best right here. With this I have decided to use a decision tree model but I am not limited to just random forest.

Before leaving the logistic regression model for good, I thought it might be beneficial to analyze which features it determined to be most important. This may offer insights elsewhere. **Figure 5** shows the feature importances for each model. These models showed some agreement with the random forest model. The opinion on effectiveness of vaccine, doctor recommendation and seasonal vaccine were all highly valued features much like random forest. The main difference is seen where the logistic regression models top 10 features contain some of the employment_industry and employment_occupation features, which the random forest model gave almost zero importance to.

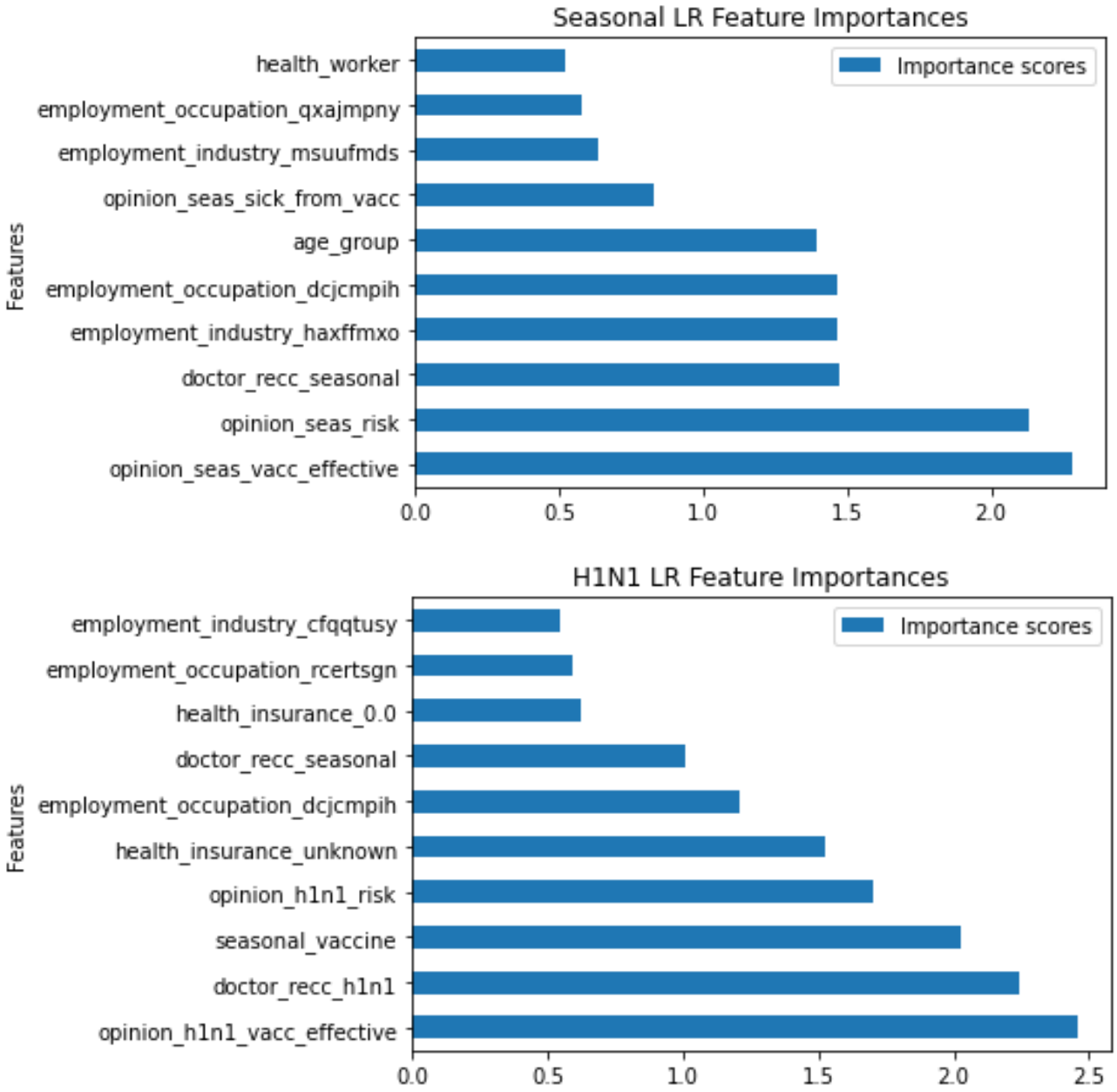


Figure 5: The 10 most important features, determined by absolute value of coefficients for each feature, for Seasonal and H1N1 logistic regression models

4.3 Gradient Boosting Classifier:

Having seen more potential in decision tree models thus far, I next looked into a gradient boosting classification model. Although similar to random forest classification, gradient boosting models have more hyperparameters worth tuning for best performance. In order to prevent overfitting, I used early stopping so that the model stops after 10 iterations are seen without increase in validation score. I set this to 10 for the seasonal vaccine as 20 seemed to overfit to the training data. This was set to 20 for the H1N1 vaccine as it included an additional variable of high importance and would therefore be more likely to withstand overfitting the training data. After setting this, I used GridSearchCV to compare the following hyperparameters: learning_rate, min_samples_leaf and max_features.

The seasonal vaccine GridSearchCV gave the following hyperparameters: learning_rate = 0.15, min_samples_leaf = 20, max_features = 'auto'. GridSearchCV for the H1N1 vaccine was relatively similar with the following hyperparameters: learning_rate = 0.15, min_samples_leaf = 10, max_features = 'auto'. It seems the seasonal model does better with a larger number of samples at each leaf which is likely due to the seasonal model being more likely to overfit the training data. With these hyperparameters the mean accuracy was 0.7840 and 0.8668 for seasonal and H1N1 respectively. The mean ROC_AUCs were 0.8607 and 0.8888 for seasonal and H1N1 respectively. The average of the two ROC_AUCs was 0.8748 which is moderately higher than the other two models.

Feature importances were also ranked for these models and gave similar results to the previous models. **Figure 6** shows a graphical representation of the top 10 features and their respective importance scores. Similar to the other models, the most important features for the seasonal vaccine were opinions on effectiveness/risk of the vaccine as well as doctor recommendation and age group. The H1N1 vaccine was similar, with doctor recommendation, opinions on effectiveness, seasonal vaccine status. One notable difference is that the H1N1 model gave more importance to health insurance status that the seasonal did not deem as important. As can be seen in the following figure, the GradientBoosting models put more importance into fewer features than the previous models. This makes me concerned about possible overfitting but I feel I put in the necessary safeguards to prevent this from occurring.

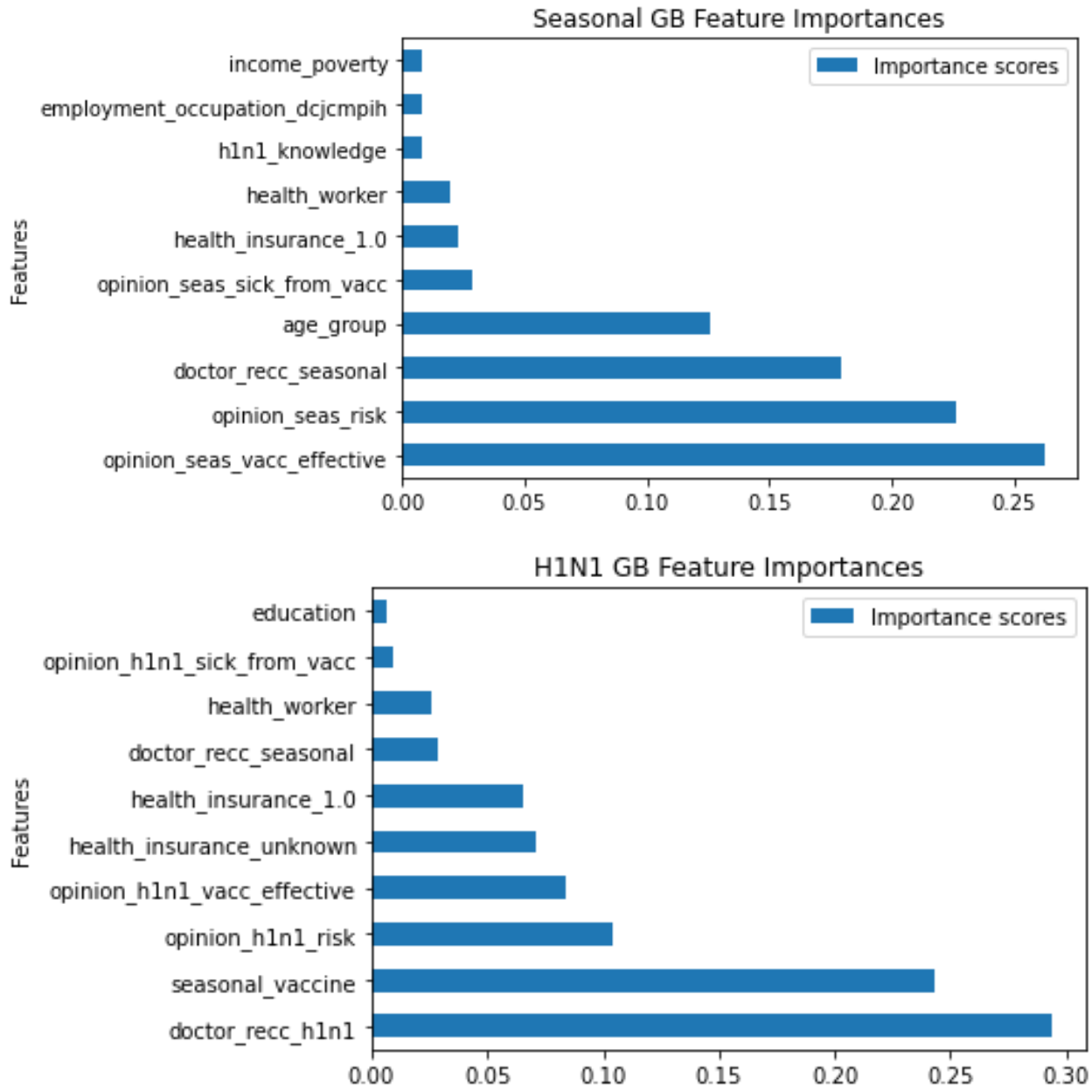


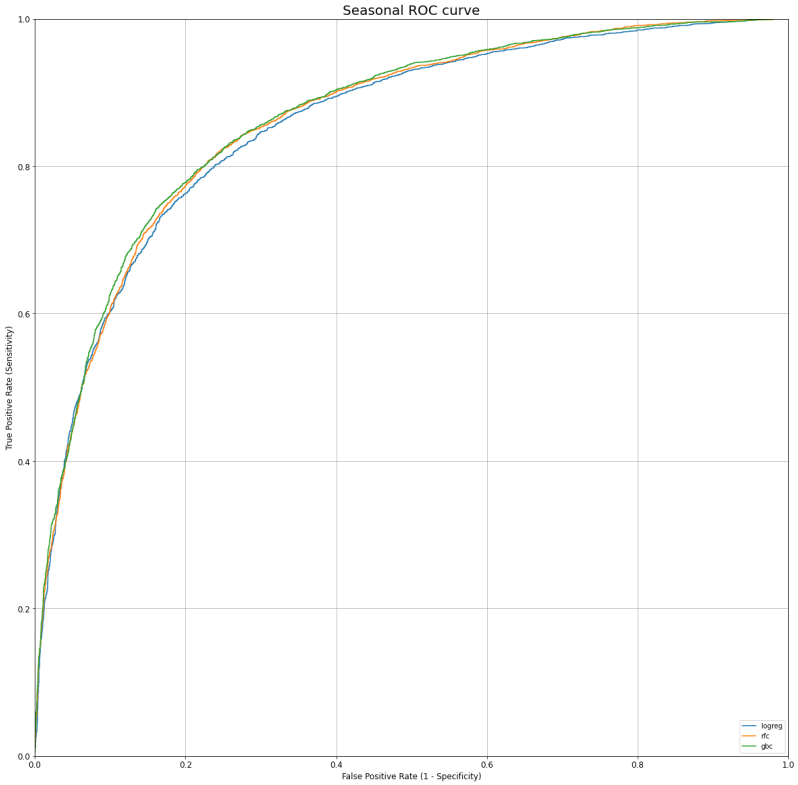
Figure 6: The 10 most important features with their scores shown for both seasonal and H1N1 GradientBoostingClassifier models

4.4 Model Selection:

Based on average accuracies and average ROC_AUC, it seems that a GradientBoostingClassifier model is the best performing model. Although all three models were all relatively successful at modeling both seasonal and novel flu vaccination status. I compared these 3 models after seeing they were the top performing. Other models I tried initially such as KNN and SVM were much worse at predicting vaccination status. I feel that for future practical purposes, a RandomForestClassifier or LogisticRegression model would work just fine in making reasonable predictions. For my current purposes of entering a contest, I want to go with the highest performing model no matter how cumbersome or wasteful it may be. For my final admission to the contest, I will be using a GradientBoostingClassifier model tuned as shown above.

When ROC_AUC was graphically represented for these 3 models in **Figure 7**, we can see that the GradientBoostingClassifier models were the highest scoring at almost all times. RandomForest only beat out Gradientboosting for a very small fraction of the H1N1 graph. Given that ROC_AUC will be used to judge the contest, this is the strongest indicator that a GradientBoostingClassifier model should be used for final prediction making. GBC happened to beat out the other models in accuracy as well as most other common metrics of performance, but just very marginally.

For any future models to actually be used by health officials, any of the 3 compared models could be used to make reasonable predictions. Depending on what was most important to the officials, each of these 3 models could be advantageous. A logistic regression model would be ideal if officials wanted a model that was very quick at predicting using a large number of features. If speed of prediction was less important and more weight was put into shortening the survey and reducing the number of features, a random forest model might be ideal. Although prediction may take a little longer, having the features reduced by at least half and models being identical for both seasonal and H1N1, may make a random forest model best. Gradient boosting models are also good for feature reduction but require more fine tuning to each model, instead of the one-size-fits-all nature of the random forest model. For my purposes, I will be using a gradient boosting model but future uses of this model may want to consider a random forest model as it performed almost identically with less fine tuning and simpler feature selection/reduction.



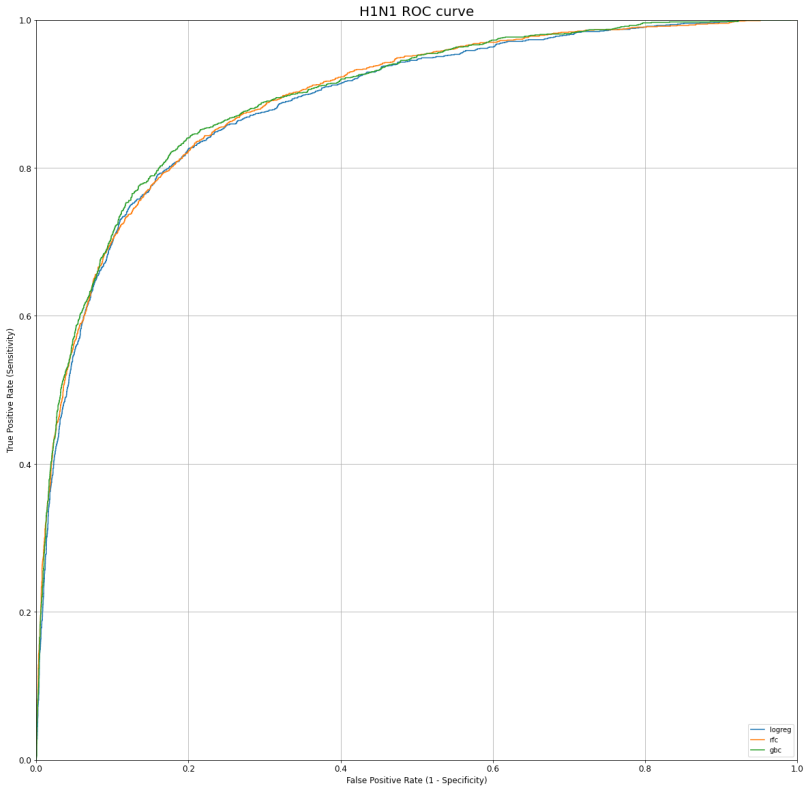


Figure 7: ROC_AUC curves for seasonal and H1N1 vaccination status