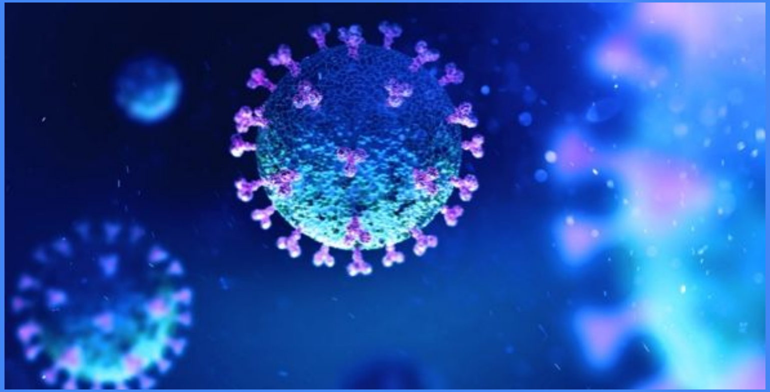
Predicting Flu Vaccinations:

Using survey questions to predict which respondents will receive a seasonal flu vaccine or novel flu vaccine

The Problem



- Hard to predict vaccination status based off demographics alone
- Spread of misinformation/conspiracies
- Not enough clean data on COVID
- Public health efforts not targeted with precision
- Herd immunity has only been a reality for a few viruses
    - Virus eradication even less common

The Data

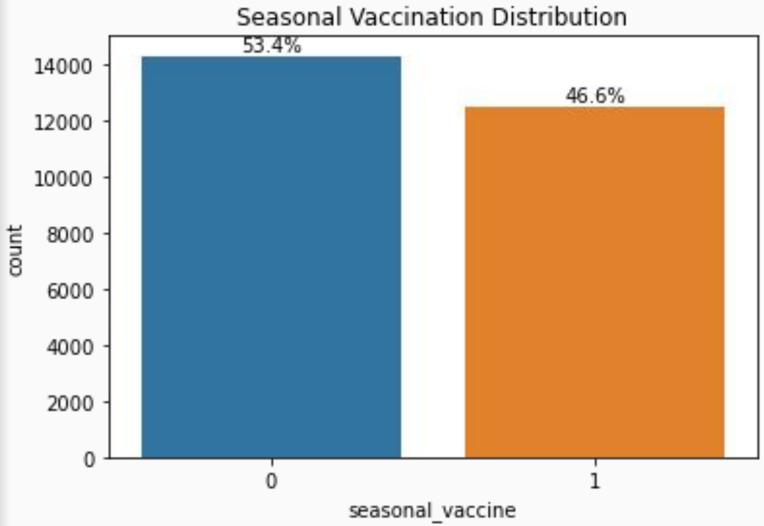
- CDC’s National Center for Health Statistics (NCHS) and National Center for Immunization and Respiratory Diseases (NCIRD)
  - The National 2009 H1N1 Flu Survey (October 2009 and June 2010)
  - 35 questions:
    - 13 yes/no questions about behaviors, health and demographics
    - 8 opinion questions rated 0-5
    - 14 demographic questions with categorical answers

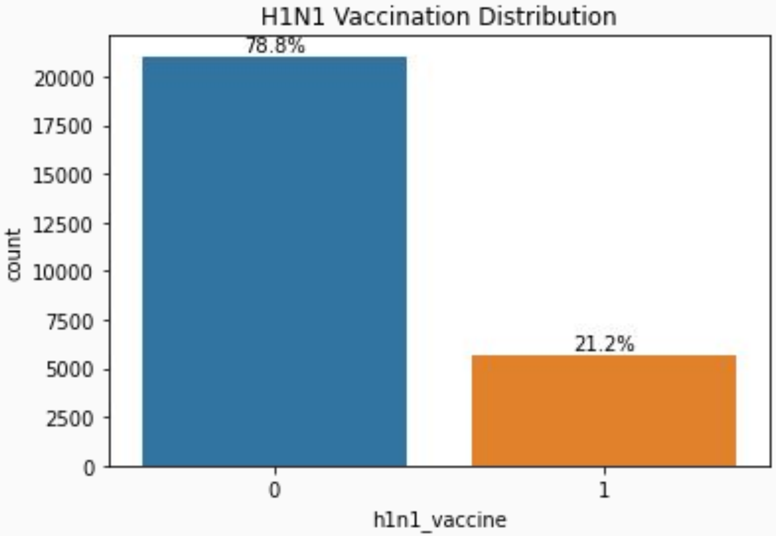
Data Cleaning

- Relatively clean raw data
- Missing values addressed by using mode of that particular column/feature
  - Non-Normal distributions
  - For all but 3 features
- 3 columns given additional categorical response of “unknown”
  - Due to large proportion of missing values

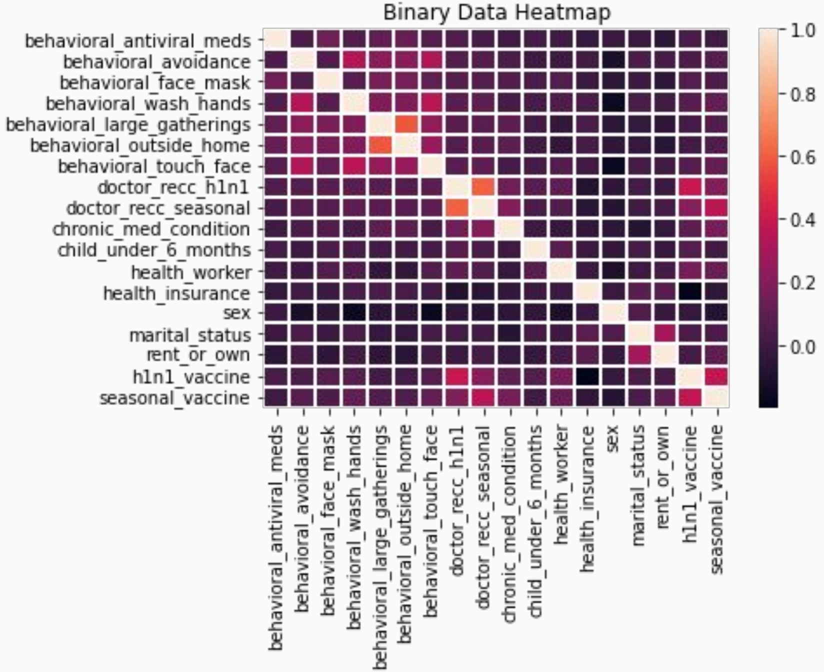
Data Exploration

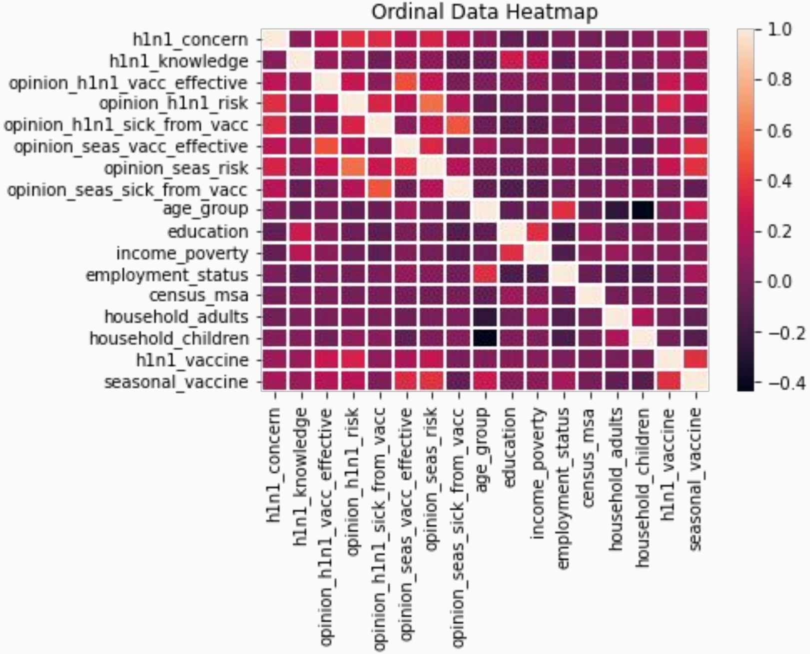
- Target variables were binary:
    - 0/1 for no/yes
    - H1N1 vaccine and
    - Seasonal flu vaccine
    - Relatively balanced
        - H1N1 less so





Data Exploration: Heatmaps

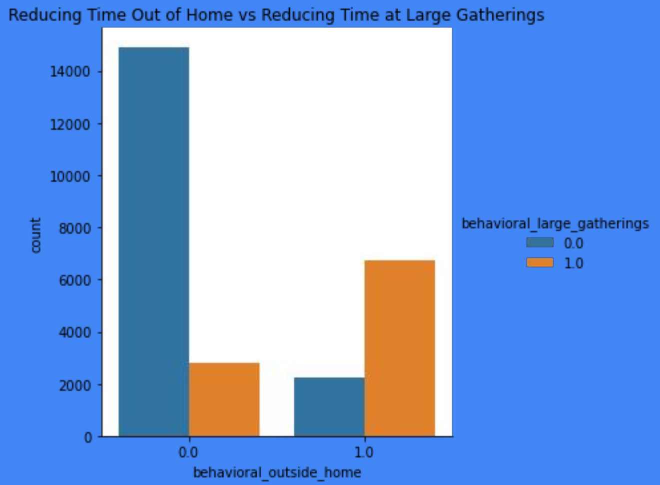


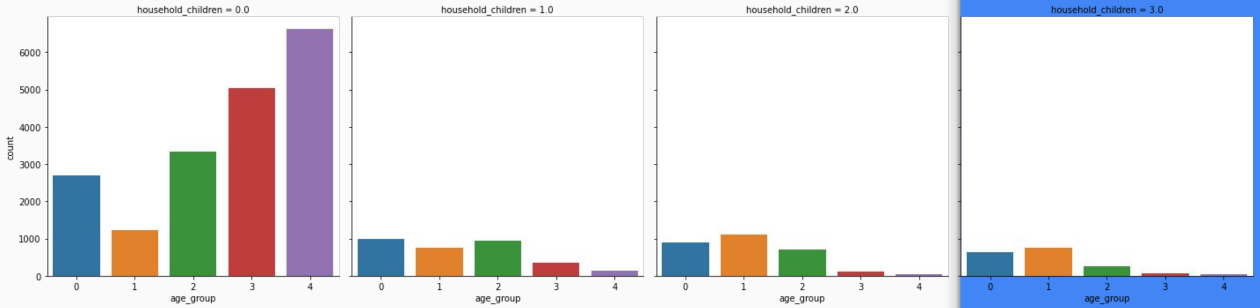


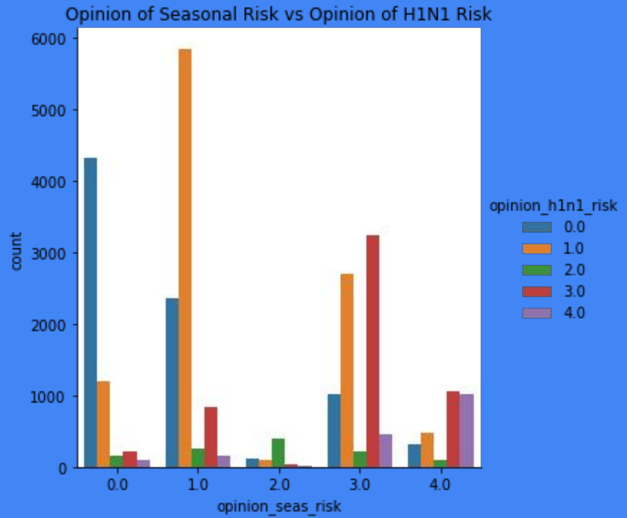
- Heatmaps of binary data and ordinal data used to visualize correlations
- Few strong correlations seen
- Good candidate for machine learning instead of simple correlation study

Strongest Correlations

- Highest 3 correlation coefficients were 0.58, 0.56 and -0.43
- Above scores for behavioral, opinion and demographic features respectively







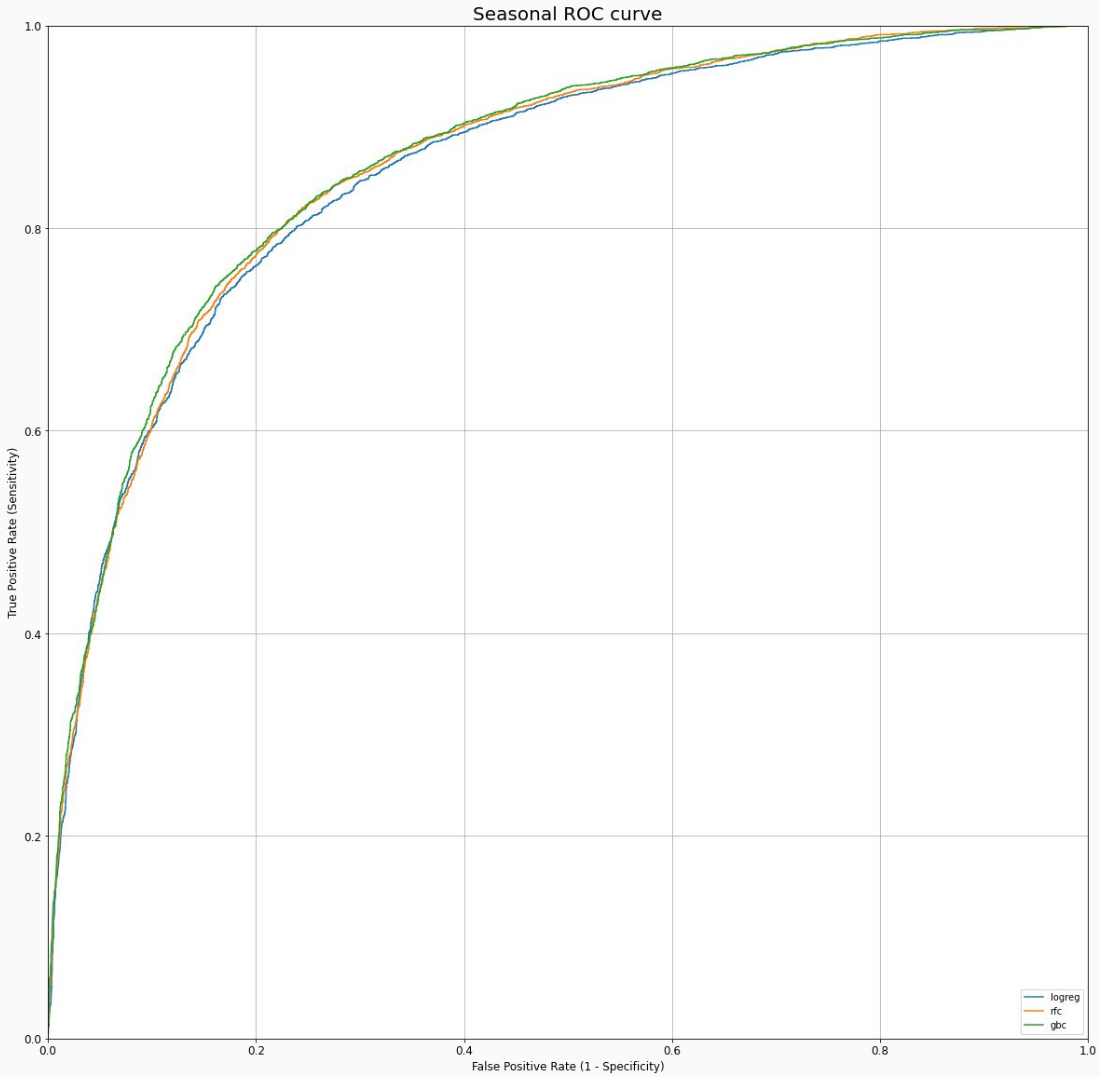
Modeling Overview/Steps

1. Data Preprocessing:
    a. One Hot Encoding of categorical variables
        i. Results in ~60 additional columns
    b. Splitting data into train and test sets
        i. 70/30 split
    c. Scaling only necessary for certain models
    d. Data is already relatively well balanced
2. Hyperparameter Tuning:
    a. Used scikit-learn’s GridSearchCV
        i. 5 fold cross validation
        ii. Evaluation metric = ROC_AUC
3. Train models on 70% of data (train set)
    a. Evaluate model performance on remaining 30% (test set)

Model Comparisons

- ROC_AUC used to choose best of 3 most accurate models
- Cross validated accuracy, roc_auc and CPU time also compared

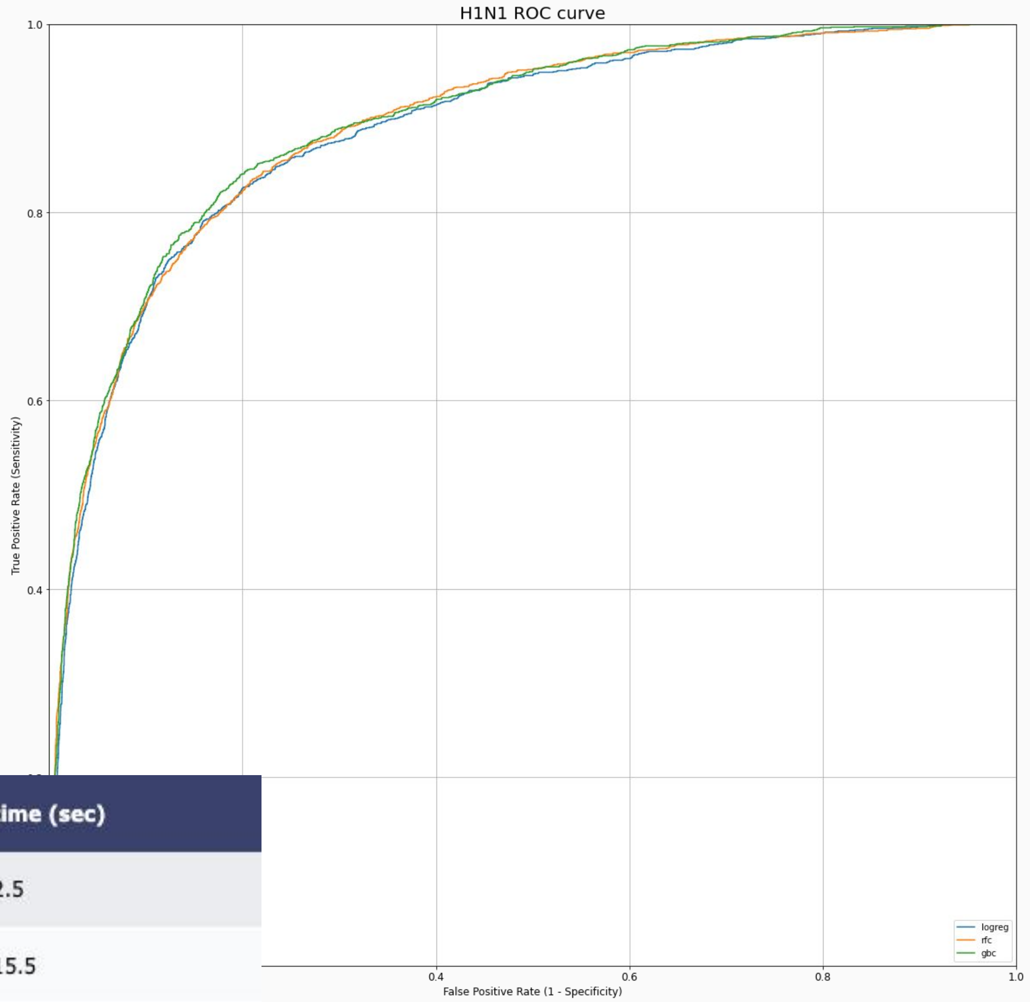
Seasonal ROC_AUC curves shown to right



Model Comparisons

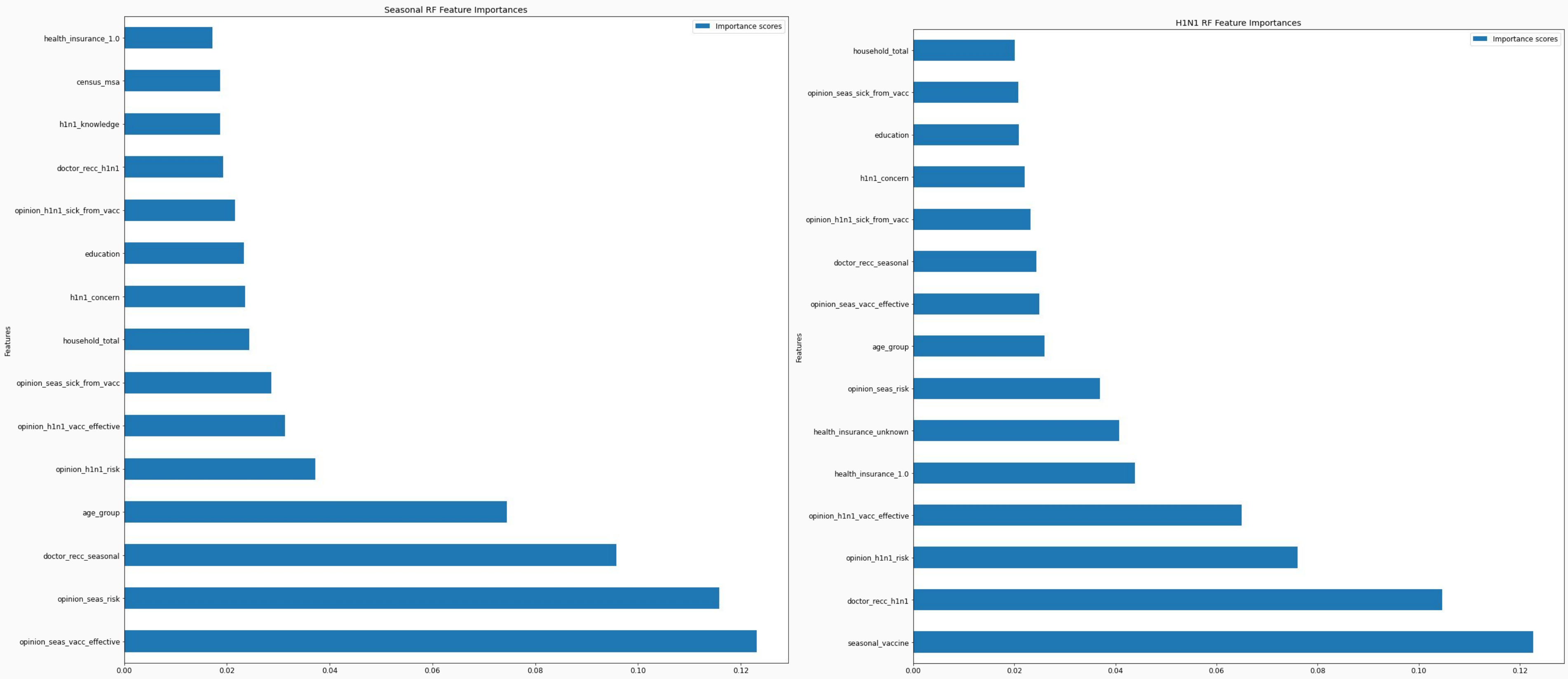
H1N1 ROC_AUC curves shown to right

Averages of seasonal and h1n1 metrics shown below

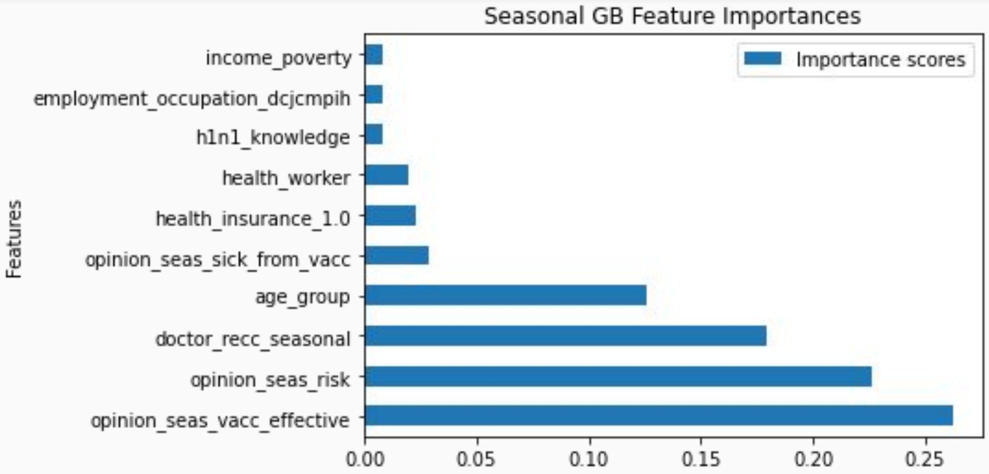


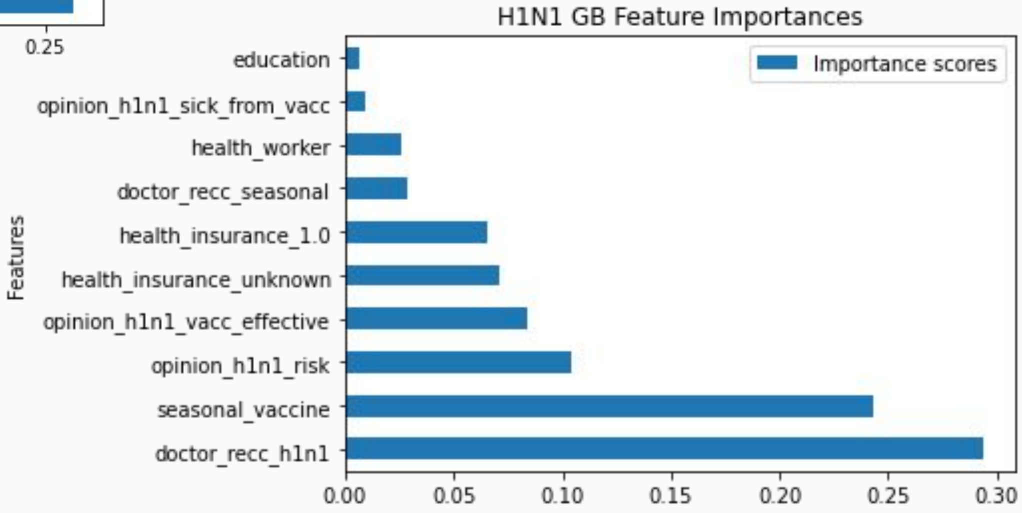
model	accuracy	ROC_AUC	time (sec)
LogisticRegression	0.821226	0.869871	2.5
RandomForestClassifier	0.822537	0.870881	15.5
GradientBoostingClassifier	0.825408	0.87478	9.0

Random Forest Feature Importances



Gradient Boosting Feature Importances





Using the GBC Model

- The GradientBoostingClassifier model was shown to be most accurate and displayed the highest ROC_AUC
- ROC_AUC was the metric of choice for the contest being entered
- First submission to contest resulted in a ROC_AUC of 0.8535 which placed 490th among 4000 entries
    - With minor tweaks I was able to increase ROC_AUC to 0.8572, placing 410th out of 4000
- My model submission scored better than 90% of all other competing user’s submissions

Conclusions and Future

- Any of these 3 models provide reasonable results
  - These 3 were compared after KNN, GNB and others were deemed least accurate
- Choice between these 3 depends on what health officials find most important
  - Lowest CPU time → Logistic Regression
  - Feature reduction → Random Forest
  - Simplest tuning → Random Forest
  - Highest accuracy/roc_auc → Gradient Boosting
  - Most complex/fine tuned → Gradient Boosting
- Future questionnaires may want to increase number of opinion questions and experience questions as opposed to demographic questions
- My Recommendation: Random Forest Classifier
  - Extremely close to GBC in accuracy but requires less tuning
  - Allows for large reduction in features/questions on questionnaire