Flu Shot Learning: Predicting H1N1 and Seasonal Flu Vaccines

December 10 2021

Gabriel Kiprono, Jarod Teague, Dipayan Banik

Introduction

Beginning in Spring 2009, the H1N1 influenza virus, generically named swine flu, swept across the world. Researchers estimated that in the first year, it was responsible for between 151,000 to 575,000 deaths globally.

A vaccine for the H1N1 flu virus became publicly available on October 2009. Later 2009 and early 2010, the US conducted a National 2009 H1N1 Flu Survey via phone call. This survey was asking participants whether they had received the H1N1 vaccine alongside other questions i.e demographic background, social, economic, opinions on risks of illness, vaccine effectiveness and behaviors towards mitigating transmission

Data

- The dataset contain the results of the survey, 36 columns. One unique respondent_id and other 35 are features.
- Predictor variables consist of binary and categorical variables
- Target variables are h1n1_vaccine and seasonal_vaccine, response was recorded as True(1) / False(0)

Objectives

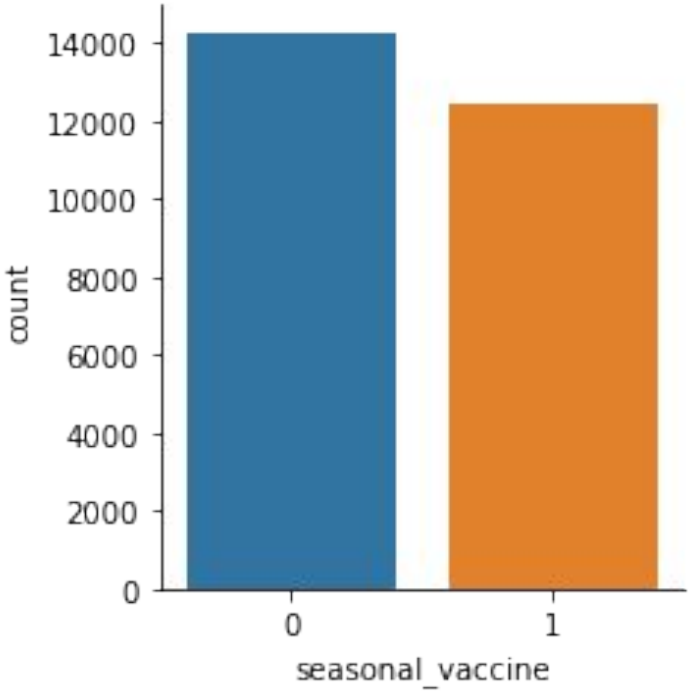
- Predict whether an individual will get H1N1 and or seasonal flu shot given information about their demographic background, opinions, health behaviours and economic status.
- Researching machine learning techniques and applying them in a data science competition
- A quality model with a reasonable AUC score

Performance Metric

- Area Under the Receiver Operating Characteristic (AUC) Curve for each target variables (seasonal and h1n1 vaccine)
- Mean score of the two target variables
- Higher value == Strong Performance

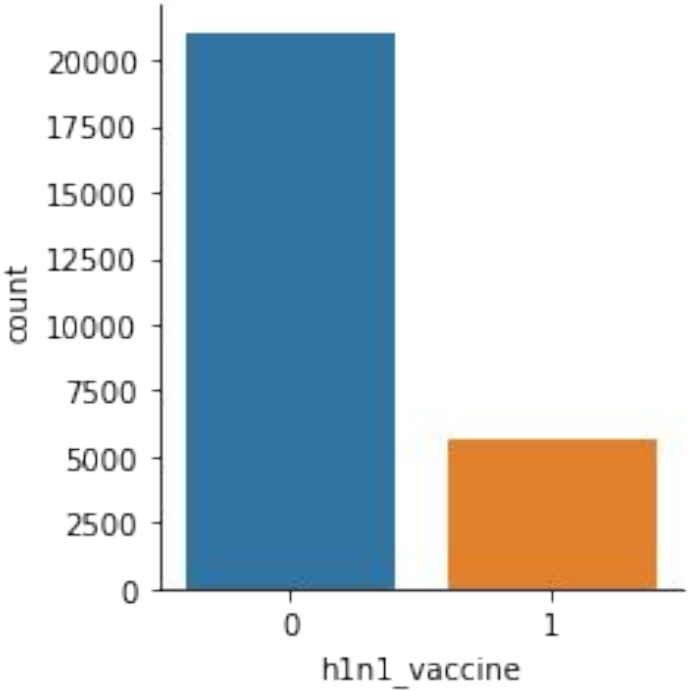
Exploratory Data Analysis

Target Variable ~ Seasonal flu



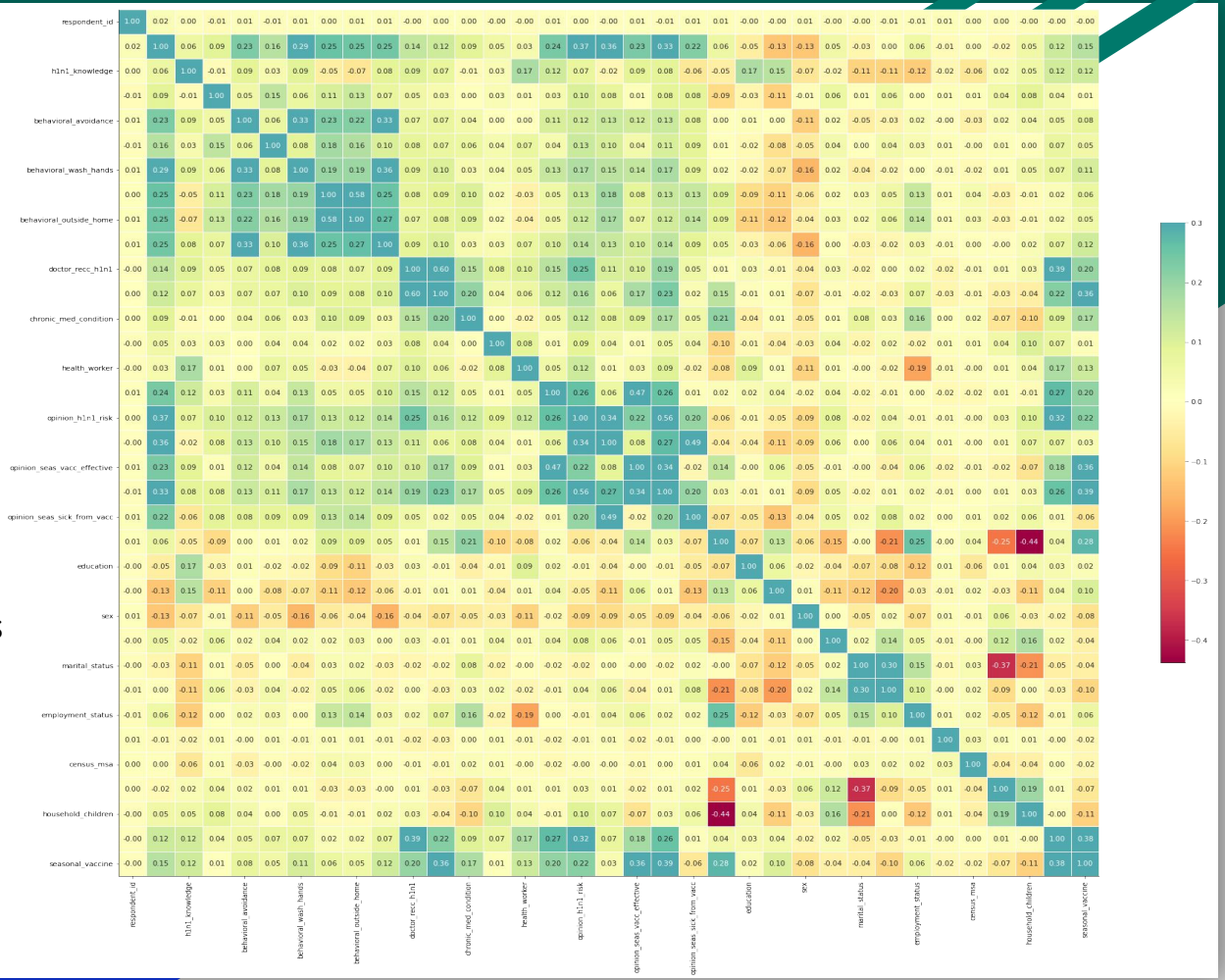
Number of people taking the seasonal vaccine is almost equal to the number of people who did not take it.

Target Variable ~ H1N1 - h1n1_vaccine

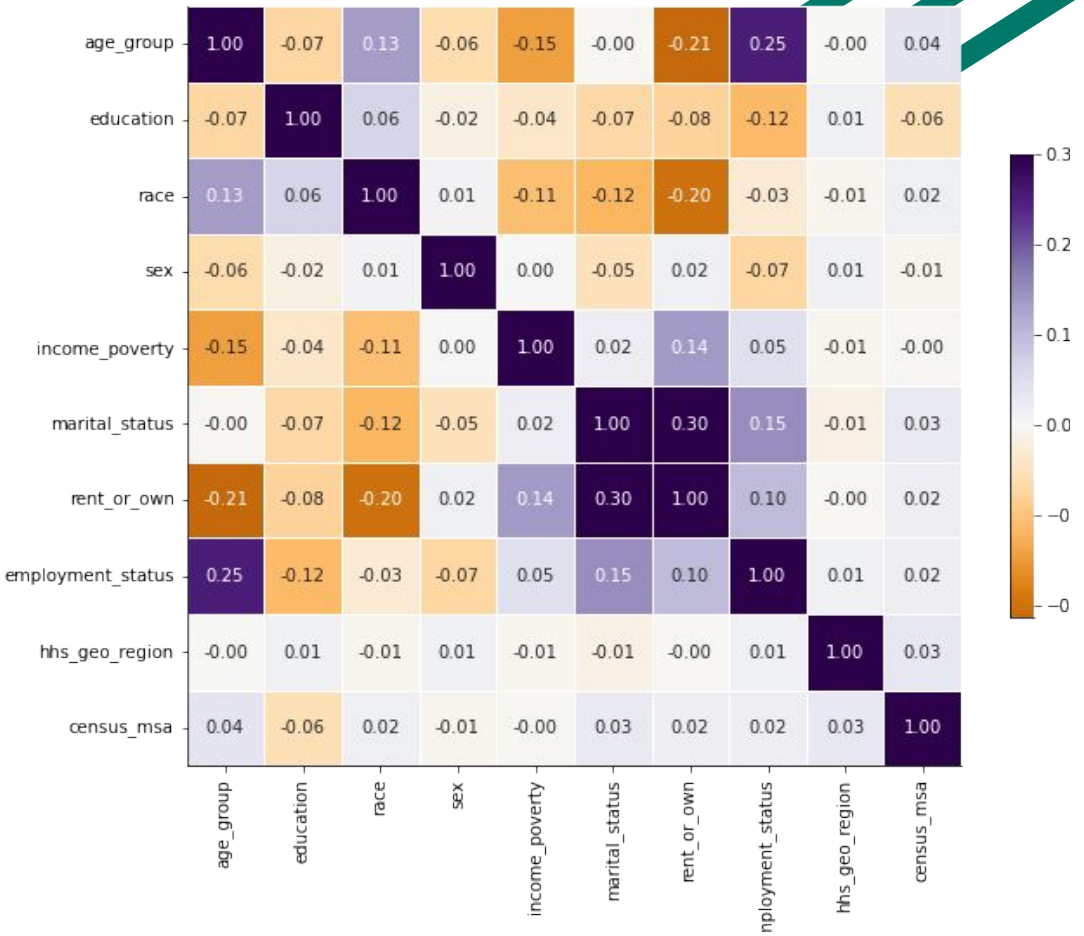


The number of people who got the H1N1 vaccine is ¼th of the surveyed population.

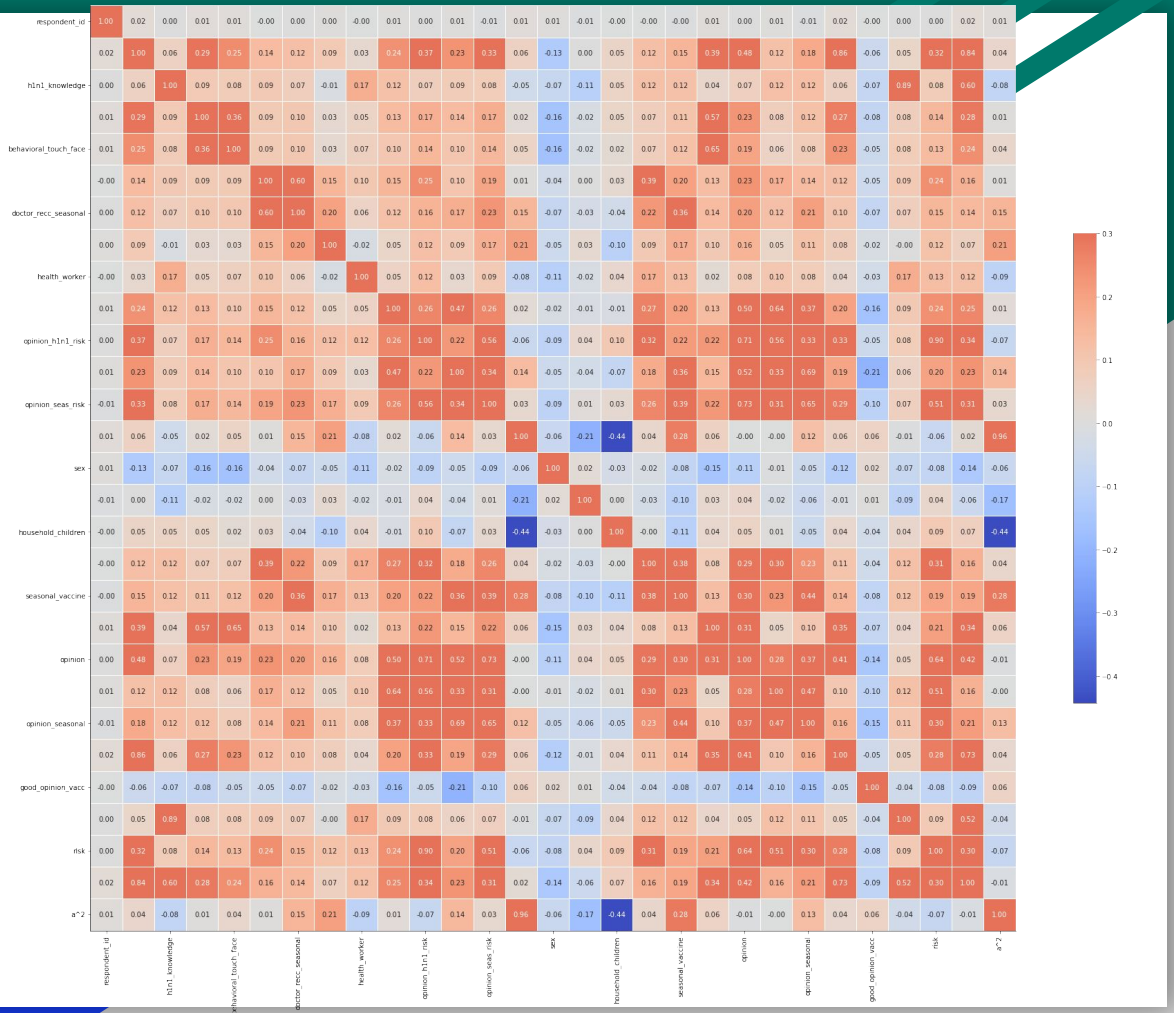
- Heatmap of the binary variables that were encoded
- High positive correlation between the behavioral features
- High positive correlation between opinion of H1N1 risk, doctors recommendation of vaccines vs if the person really takes the vaccine.



- Heatmap of the categorical features that were encoded
- Lots of negative correlation between some of the features



- Heatmap of both the categorical and binary variables
- Overall, a positive correlation with a lot of red color squares



Feature Preprocessing and Data Preparation

- Numeric columns processing
  - Scaling - Standard Scaler
  - NA Imputation - Simple Imputer
- Non-numeric columns processing
  - One hot encoder
- Column Transformer

- Train and Test set
  - 70-30, 60-40
- Train Model and Evaluate Performance
- Feature Scaling

Modeling

Models tried so far:

- Naive Bayes
- Logistic Regression
- Decision Trees
- Random Forest
- K-Nearest Neighbors
- SVM
- XGBoost

Naive Bayes

- “Naive” approach - easier to implement
- Takes advantage of Bayes Theorem - Conditional Probability
- Disadvantage: Assumption of independent predictors

K-Nearest Neighbors (KNN)

Birds of the same feathers flock together

- Supervised MLA ~ relies on labelled input
- Solve both classification and regression problems
- Simple product/movie/article recommendation system
- Slows down with increase in volume of data
- N -number?

Logistic Regression

- Lasso (L1) regularization for feature selection
- Zeroes out some coefficients
- Disadvantage: Linearity between the variables

Decision Tree

- Flow chart method - if, else statements
- Easier to generate rules
- Disadvantage: Unstable - might perform too well with the training data

Random Forest

- Ensemble learning method
- Combines multiple decision trees so accuracy is higher
- Disadvantage: Overfitting

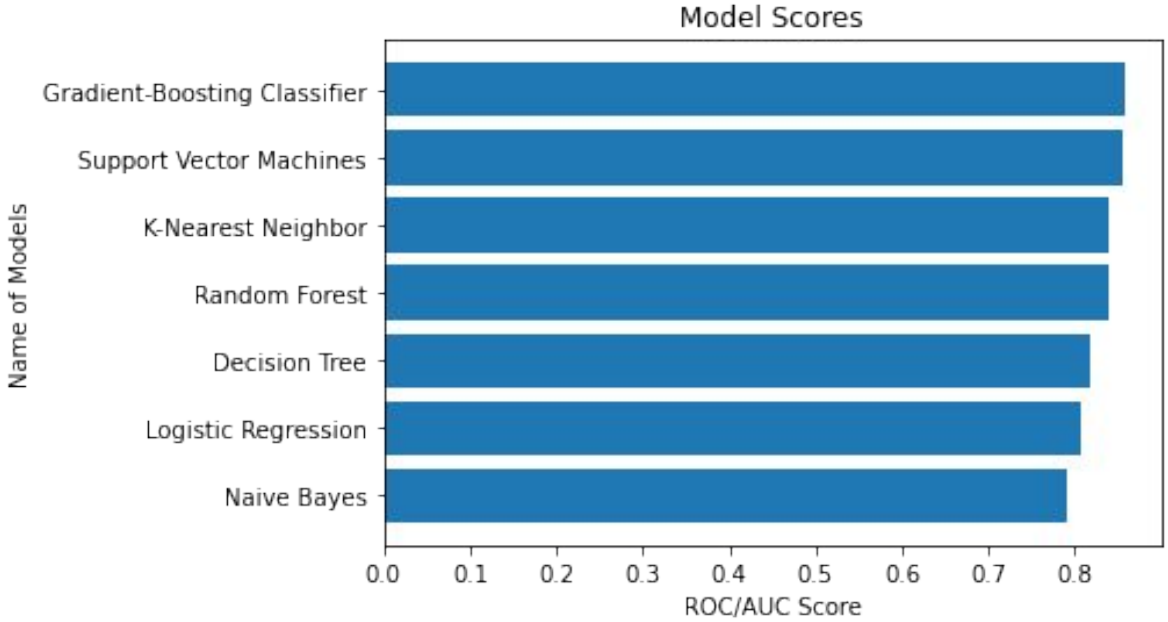
Support-Vector Machine (SVM)

- Supervised learning model
- “Maximizes” margins
- Kernel trick
- Disadvantage: Required training time is higher when the dataset is big

XGBoost

- Extreme Gradient Boosting
- Boosting is an ensemble technique
- Uses gradient descent to minimize the loss when adding new models
- Supports both regression and classification
- Disadvantage: Sensitive to outliers

Performance Metric - Results



Submission Results

Submissions

BEST	CURRENT RANK	# COMPETITORS	SUBS. MADE
0.8499	433	3259	2 of 3

SUBMISSION RESTRICTIONS

Setbacks

- Grid Search CV
- Hyperparameter tuning
- Deep learning - Neural Networks

Takeaways

- Team Collaboration
- Data Science Project Experience
- Research Experience
- Articles, Sources
- Communications

Questions?