**Introduction**

With the fast spread of Covid-19, people are widely recommended to receive covid vaccination to reduce the infection rate. Some people, however, hold their own perspectives and choose to not be vaccinated. At this report, we aim to build a predictive model to determine the potential demographic and social-economic factors that affect U.S. citizens’ vaccination status and predict their vaccination behavior.

To achieve the goal, we fitted multiple models to determine which variables affect people’s Covid vaccination status, and chose an optimal model based on model comparison. The dataset contains 19 variables and 8308 observations. The response variable is covid\_vaccination, which indicates whether a person receives their Covid-19 vaccination. There are 18 predictors including:

* id (member ID)
* cons\_chmi (census median household income)
* est\_age (member age)
* hum\_region (member geographic information)
* atlas\_percapitainc (per capita income in the past 12 months 2014-2018)
* rwjf\_resident\_seg\_black\_inx (social and economic factors - residential segregation - black/white)
* rwjf\_uninsured\_adults\_pct (clinical care - percentage of adults under age 65 without health insurance)
* atlas\_hh65plusalonepct (percent of persons 65 or older living alone)
* atlas\_medhhinc (median household income)
* cons\_lwcm07 (the probability of the individual being less likely to use doctor/physician as a primary source for medical information)
* atlas\_pct\_sbp15 (School Breakfast Program participants (% pop))
* atlas\_povertyallagespct (poverty rate)
* cons\_rxadhm (rx adherence – maintenance)
* race\_cd (Code indicating a member's race {0 = Unknown, 1 = White, 2 = Black, 3 = Other, 4 = Asian, 5 = Hispanic, 6 = N. American Native})
* atlas\_low\_education\_2015\_update (low education counties)
* atlas\_type\_2015\_mining\_no (mining-dependent counties)
* lang\_spoken\_cd (preferred language for member)
* sex\_cd (member gender)

With the dataset and data modeling, we are trying to answer the following questions:

1. What variables affect people’s Covid-19 vaccination status the most?
2. What models can be used to predict the result?
3. Which model is ultimately selected and why so?

To prepare and clean the data, we removed the ID from the variables. In addition, we removed categorical variables to graph feature plots. we split the dataset into two parts: training data (70%) and test data (30%). we set all variables except the response variable as X, and the response variable as Y. To better fit X and Y in models, we converted X into a matrix when creating training and test data.

**Exploratory analysis/visualization**

Based on the feature plots (Figure 1), we can see that the distributions of vacc and no\_vacc responses are very close to each other. Among the distribution of all variables, distributions of predictors atlas\_hh65plus-alonepct (percent of persons 65 or older living alone), rwjf\_resident\_seg\_black\_inx (black/white) are normal distributed; distribution of predictor est\_age (member age) is left-skewed. The distribution of all other predictors is right-skewed.

Graphical user interface, diagram

Description automatically generated

Figure 1 Density distribution of continuous variables in two response classes

**Models**

Since the response of this dataset only contains two classifications, we decided to fit data into penalized logistic regression (GLMNET), generalized additive mode (GAM), linear discriminants analysis (LDA), tree-based methods (Random Forest), Support Vector Machines (SVM), and Neural Networks models.

1. **GLMNET**

To address the problem of sparse data and too many predictors, we apply different strengths of L1 and L2 penalty on the Maximum Likelihood Estimation Process (Elastic Net) to improve the model. A combination of different alpha and lambda values are applied in the model, and the tuning parameters resulting in the largest ROC value are selected as the final model. As a result, the model with alpha = 0.05 and lambda = 0.076 is selected for final prediction.

1. **GAM**

To adopt the nonlinearities of variables but retain the additive structure of linear models, we applied a generalized additive model (GAM) to further increase the model flexibility. Non-linear functions are applied to each variable and the non-linearity level is automated determined during training. When we fit data into a GAM model, we deleted categorical variables as categorical variables are less tolerated in the GAM model. From the summary of the GAM model, we can see that the model uses logit link functions and assumes a binomial distribution of errors. We can also see that the model converted est\_age, cons\_chmi, atlas\_pct\_sbp15, atlas\_povertyallagespct, cons\_lwcm07, atlas\_percapitainc, atlas\_medhhinc, rwjf\_resident\_seg\_black\_inx, atlas\_hh65plusalonepct, and rwjf\_uninsured\_adults\_pct predictors. The model didn’t convert atlas\_low\_educa-tion\_2015\_update, race\_cd, and cons\_rxadhm since these predictors are not linear (Figure 2).

Text, letter

Description automatically generated

Figure 2

1. **LDA**

In addition to the linear models, we also performed Linear Discriminant Analysis (LDA). Since the LDA model only accepts numeric variables, the three categorical variables other than the outcome this project researched are all omitted in consideration of the predictor consistency in the later model comparison. Eventually, there are 14 predictors included in these models. Based on the LDA plot, the two categories of covid vaccination show very similar distributions.

1. **Random Forest**
2. **Support Vector Machine**
3. **Neural Network**

**Model Comparison**

When comparing these six models, we graphed a box plot of six models’ ROC. From figure ? we can see that the Random Forest model has the largest ROC value. GLMNET has the second largest ROC value, followed by GAM, LDA, linear SVM, and radial SVM has the smallest ROC value. Therefore, we can conclude that Random Forest has the optimal value based on the cross-validation of training data.

Chart, box and whisker chart

Description automatically generated

Figure ?

For the test data performance, we compared the AUC of six models. From figure ? we can see that the AUC values of Random Forest, GLMNET, LDA, GAM, linear SVM, radial SVM, and Neural Network models are 0.592, 0.59, 0.586, 0.575, 0.559, 0.532, and 0.516, respectively. Therefore, Random Forest is the optimal model at test data performance.

Chart, line chart

Description automatically generated

Figure ?