In this project, we fitted multiple models to predict whether people’s vaccination status depends on the dataset, and ultimately, we chose the 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.

Based on the feature plots, 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.