BISTP8106 Data Science II
Midterm Project Report
Jiaqi Chen (jc5681)

In this project, I developed multiple models to predict whether people's vaccination status depends on the dataset, and 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:

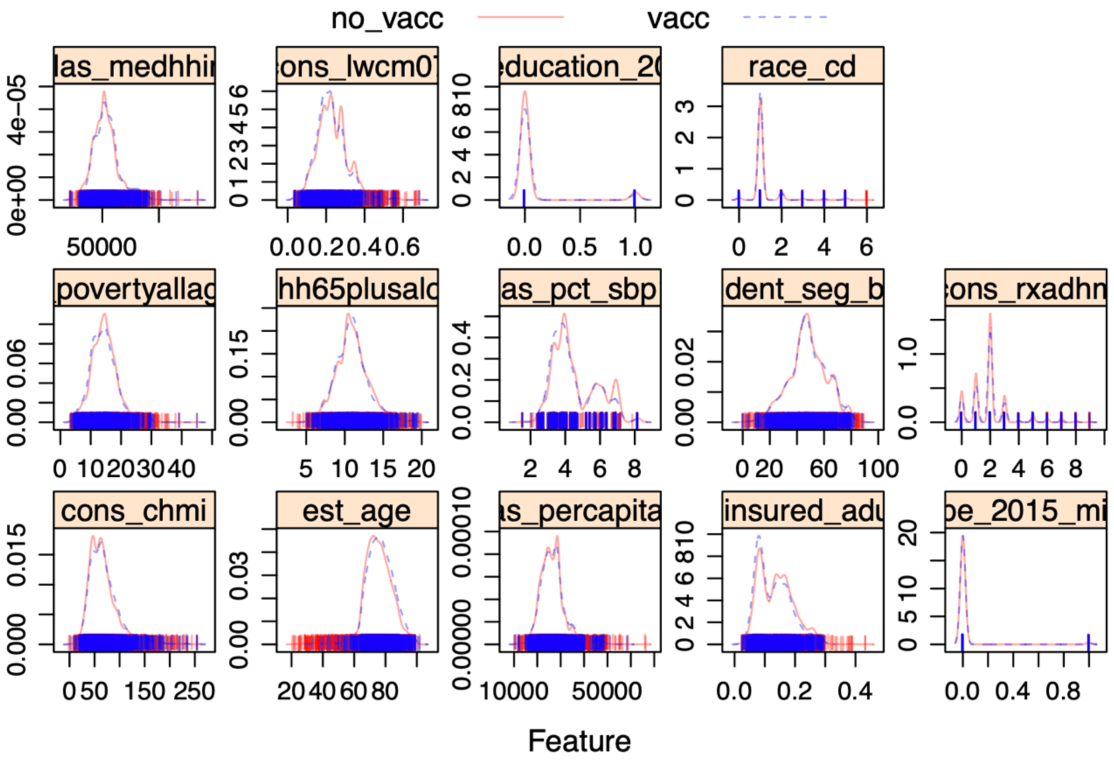
- id (member ID)
- cons_chmi (census median household income)
- est_age (member age)
- hum_region (member geographic information)
- atlas_percapitainc (per capita income in the past 12 months 2014-2018)
- rwjf_resident_seg_black_inx (social and economic factors - residential segregation - black/white)
- 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, I am 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?

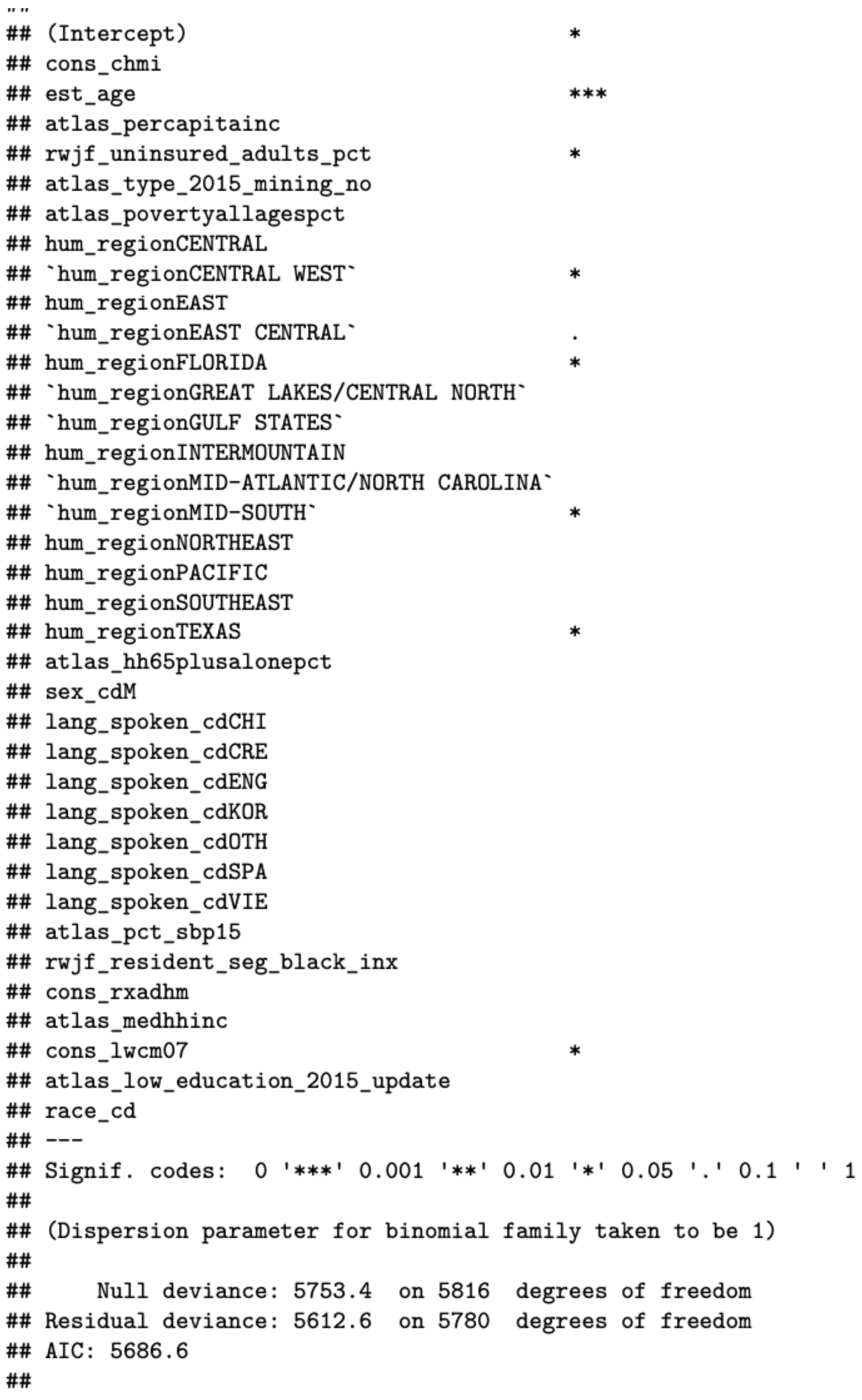
To prepare and clean the data, I removed the ID from the variables. I also removed categorical variables to graph a feature plot. I split the dataset into two parts: training data (70%) and test data (30%). I set all variables except the response variable as X, and the response variable as Y. To better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

Based on the feature plot, 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 are right-skewed.



Plot 1

Since the response of this dataset only contains two classifications, I decided to use logistic regression to fit models and analyze data. More specifically, I fitted data into GLM, GLMN, GAM, and MARS models. From the summary of the GLM model, we can see that predictors est_age (member age), rwjf_uninsured_adults_pct (percentage of adults under age 65 without health insurance, hum_region (Member geographic information), and cons_lwcm07 (the probability of the individual being less likely to use doctor/physician as a primary source for medical information) are statistically important variables as their p-values are less than 0.05 (Plot 2). Since the residual deviance of the GLM model is 5612.6, closer to 5780 degrees of freedom, the GLM model is proven as a good fit. With the confusion matrix of the GLM model, we can see that the model accuracy is 0.8045 with (0.7884, 0.8199) 95% confidence interval and $2e^{-16}$ P-value. Therefore, the GLM model is considered accurate.



```
""
## (Intercept)                              *
## cons_chmi
## est_age                                  ***
## atlas_percapitainc
## rwjf_uninsured_adults_pct                *
## atlas_type_2015_mining_no
## atlas_povertyallagespct
## hum_regionCENTRAL
## `hum_regionCENTRAL WEST`                 *
## hum_regionEAST
## `hum_regionEAST CENTRAL`                 .
## hum_regionFLORIDA                        *
## `hum_regionGREAT LAKES/CENTRAL NORTH`
## `hum_regionGULF STATES`
## hum_regionINTERMOUNTAIN
## `hum_regionMID-ATLANTIC/NORTH CAROLINA`
## `hum_regionMID-SOUTH`                    *
## hum_regionNORTHEAST
## hum_regionPACIFIC
## hum_regionSOUTHEAST
## hum_regionTEXAS                          *
## atlas_hh65plusalonepct
## sex_cdM
## lang_spoken_cdCHI
## lang_spoken_cdCRE
## lang_spoken_cdENG
## lang_spoken_cdKOR
## lang_spoken_cdOTH
## lang_spoken_cdSPA
## lang_spoken_cdVIE
## atlas_pct_sbp15
## rwjf_resident_seg_black_inx
## cons_rxadhm
## atlas_medhhinc
## cons_lwcm07                              *
## atlas_low_education_2015_update
## race_cd
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5753.4  on 5816  degrees of freedom
## Residual deviance: 5612.6  on 5780  degrees of freedom
## AIC: 5686.6
##
```

Plot 2

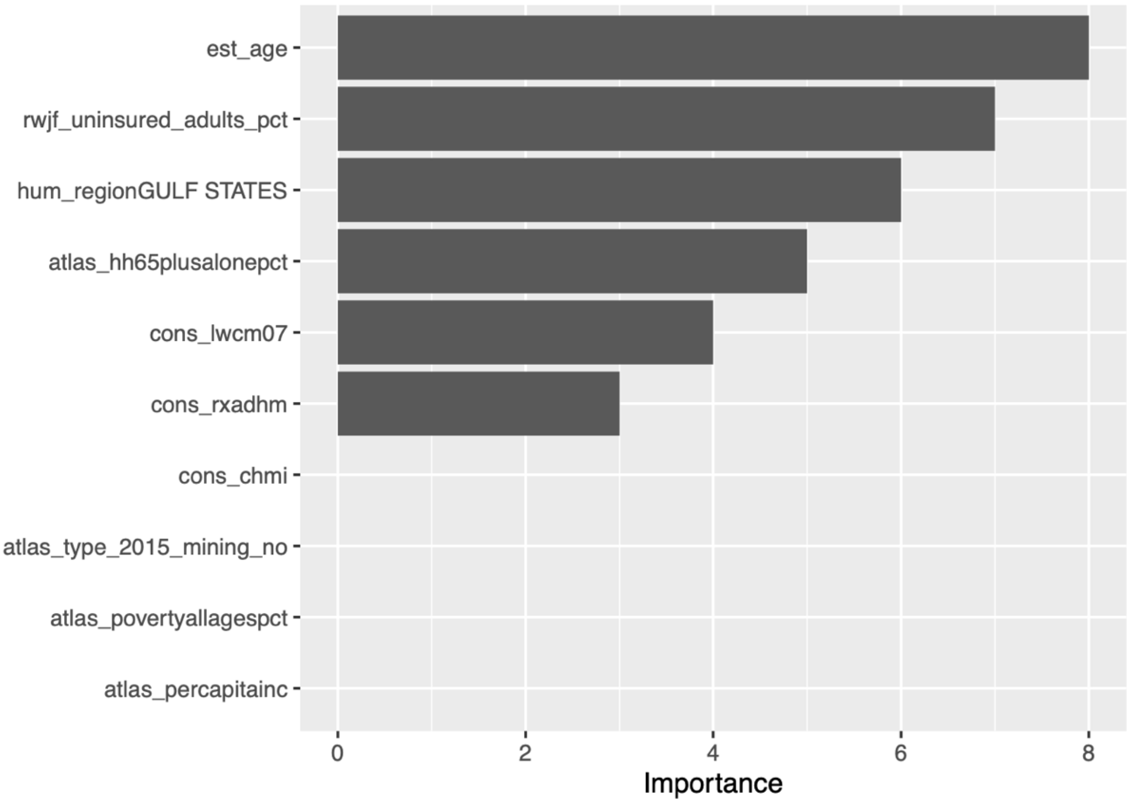
In addition, I used GLMNET model to penalize logistic regression. In the GLMNET model, I picked tuning parameters by choosing the largest alpha and lambda values. The chosen alpha and lambda values are 0.05 and 0.066, respectively.

When I fit data into a GAM model, I 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 (Plot 3).

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## .outcome ~ atlas_low_education_2015_update + race_cd + cons_rxadhm +
##     s(est_age) + s(cons_chmi) + s(atlas_pct_sbp15) + s(atlas_povertyallagespct) +
##     s(cons_lwcm07) + s(atlas_percapitainc) + s(atlas_medhhinc) +
##     s(rwjf_resident_seg_black_inx) + s(atlas_hh65plusalonepct) +
##     s(rwjf_uninsured_adults_pct)
##
## Estimated degrees of freedom:
## 2.59 1.00 5.64 1.86 1.00 3.51 1.70
## 1.00 7.34 1.15  total = 30.79
##
## UBRE score: -0.02549095
```

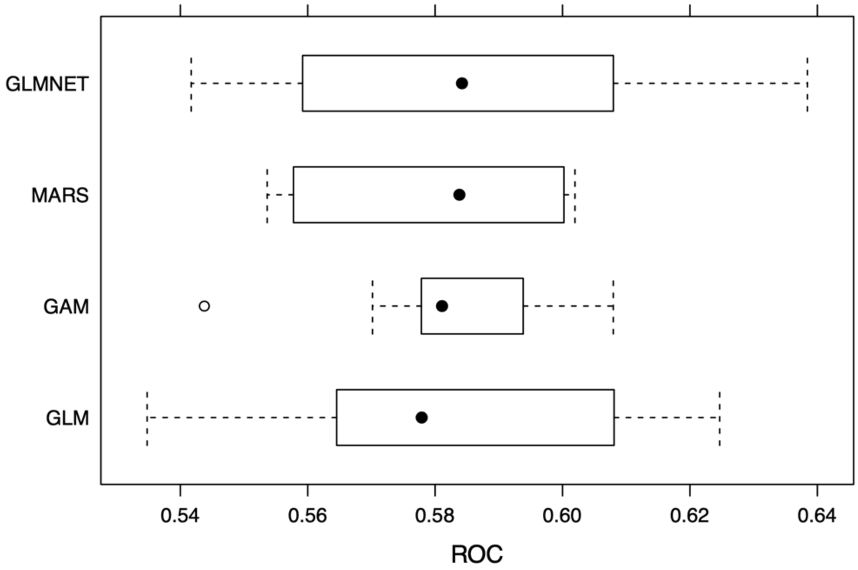
Plot 3

The last model I used is MARS. From plot 4 we can see that the tuning parameter is selected as 1 product degree at 9 number of terms. Plot 5 shows that in MARS model, est_age, rwjf_uninsured_ad-ults_pct, hum_region, atlas_hh65plusalonepct, cons_lwc-m07, and cons_rxadhm variables are important in predicting the response.



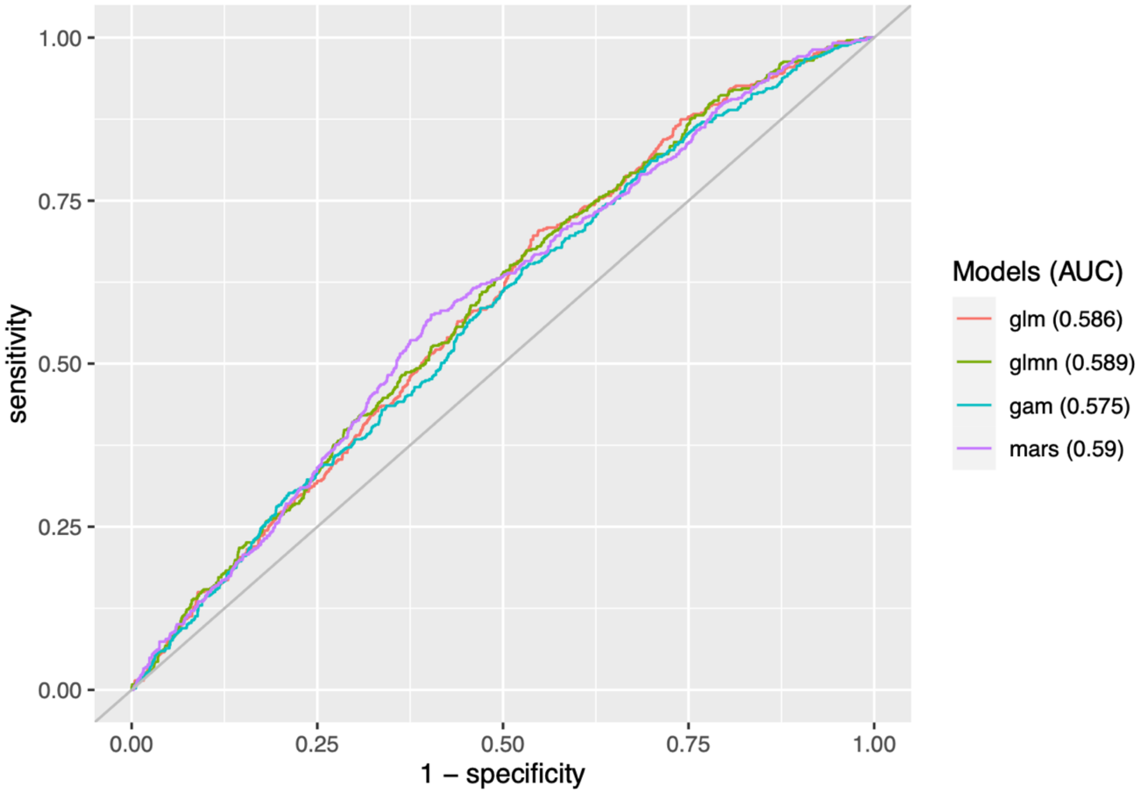
Plot 4

When comparing these four models, I graphed a box plot of four models' ROC. From plot 5 we can see that the GLMNET model has the largest ROC value. Therefore, we can conclude that GLMNET is the optimal value based on the cross-validation of training data.



Plot 5

For the test data performance, I compared the AUC of four models. From plot 6 we can see that the AUC values of MARS, GLMNET, GLM, and GAM models are 0.59, 0.589, 0.586, and 0.575, respectively. Therefore, MARS is the optimal model at test data performance.



Plot 6

Since we value more on the cross-validation of training data than test data performance, I would choose the GLMNET model to predict the data response. Compared to other models, the GLMNET model contains a limitation. More specifically, the GLMNET model can only use linear variables to predict the response variable, while the GAM model can use both linear and nonlinear variables to predict. Overall, among the predictive variables, est_age, rwjf_uninsured_adults_pct, hum_region, cons_lwcm07, atlas_hh65plusalonepct, and cons_rxadhm play important roles in predicting the response.