BISTP8106 Data Science II Final Project Report
Yiru Gong (yg2832); Yiwen Zhao (yz4187); Jiaqi Chen (jc5681)

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 (Appendix A: Figure 1), we can see 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. According to the correlation plot (Appendix A: Figure 2), there are stronger nagative correlation between member age and the probability of the individual being less likely to use doctor/physician as a primary source for medical information, and the probability of the individual being less likely to use doctor/physician as a primary source for medical information is positively correlated with rx adherence.

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.

I. 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.

II. 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 (Appendix A: Figure 3), 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).

III. 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.

IV. Random Forest

To better adapt both categorical and numerical variables, we applied a flexible tree-based model to the data, namely random forest (RF). The number of random selection of predictors in each split (mtry), and the minimum node size (min.node.size) are tuned to identify the best RF model. Gini index is applied as split criteria for the categorical response Y, and ROC is used as evaluation metrics. As a result, mtry = 2 and min.node.size = 4.

V. Support Vector Machine

By considering that it is a two-class classification problem, which has responses separated in vacc and no_vacc, we use the support vector machine to see how the featured space is separated by these two classes and find out the optimal boundary of possible outputs. Due to the restrict of support vector machine, we include only numerical variables to train this model. By implementing linear and radial bases function of support vector machine from kernel lab, we found out that there is no obvious difference on the accuracy of each cost. The values of accuracy turn out to stabilize at 0.8, which makes us hard to choose the largest accuracy from the plot. In consideration of the subsequent comparison between models, we decide to try the ROC metric in these two models. According to the two plots below (Appendix A: Figure 4&5), we can clearly define that the highest ROC value which would be the tuning parameters for the final model. Moreover, by selecting the best tune from these two models, we got the tuning parameter specified for both smvl and svmr.

VI. Neural Network

To accommodate the high similarity in variables distribution in two groups, we apply a more complicated black-box model of Neural Network to the model. We set three dense layers with batch normalization and random dropouts after each layer. The layer units and dropout

probabilities are tuned to identify the model with highest accuracy (Appendix A: Figure 6). For each model, 30 epochs are set to train the model and Categorical Cross Entropy loss function is applied for optimization. The model learning rate is fitted as 0.001. As a result, we got 64 units for layer 1, 64 units for layer 2, and 128 units for layer 3. The corresponding dropout probability is 0.4, 0.2, 0.3.

Model Comparison

When comparing these six models, we graphed a box plot of six models' ROC. From figure 7 in the Appendix A, 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. For the test data performance, we compared the AUC of six models. According to the ROC curve (Appendix A: Figure 8), 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.

According to the figure 9 in the Appendix A, median household income (atlas_medhhinc) is the most important variable that affect people's decisions on get vaccinated or not, and the preferred language for member (lang_spoken_cd) may be the variable that make the least influences on people's decision on vaccination. This bar plot shows the average takes from multiple times of permutation, which makes the results turn out more stable.

Conclusion

In conclusion, the best prediction model of train data is the Random Forest model since its AUC value is highest one in these six models. By testing the models' predictive ability with the test data, which is 30% of the whole database, the Random Forest model is also the best model among all. Although random forest is a complex model which needs longer time to training, it can provide us stable algorithms and less noise. It reduces the variance to improve the accuracy as well. After comparison, the results analyzed from the train data and test data are consistent. Based on the previous model interpretation, the minimal node size selected by Random Forest final model is 4 with mtry = 2. we extract the variable importance from the final model to find out the global interpretation of the model. As the bar plot shows, the most important variable in the Random Forest model is the median household income and the least important one is the preferred language for member. Thus, since the median household income affect people's decision on taking covid-19 vaccination the most, if we want to make the vaccination rate higher in the whole population, we may consider how people's economic issues affect their decision and what kind of benefits may motivate the population to get vaccinated.

Appendix A

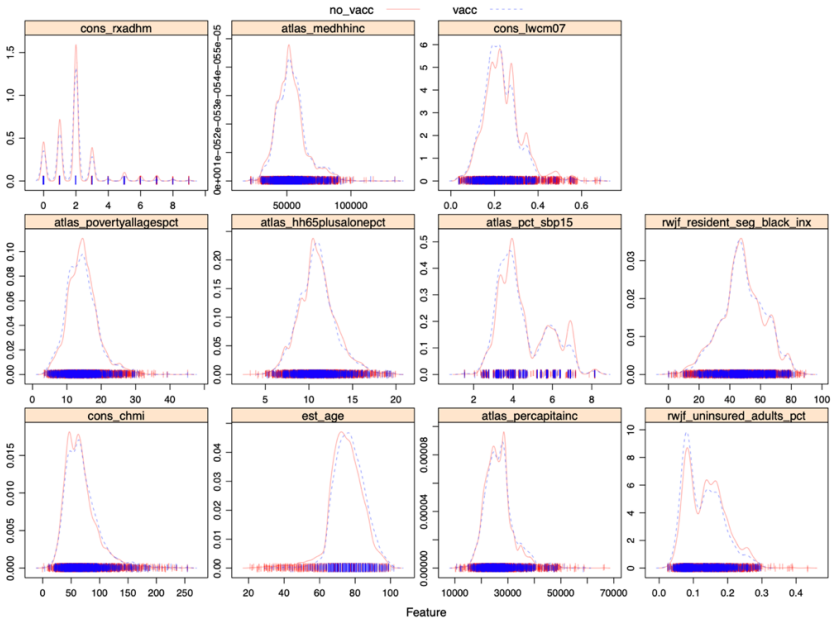

Figure 1 *Feature Plot: Density distribution of continuous variables in two response classes.*

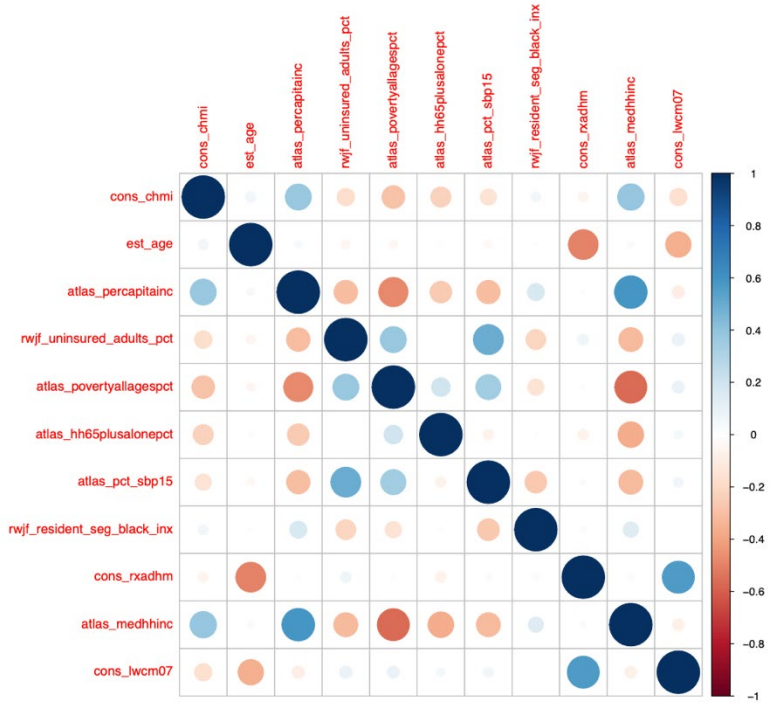

Figure 2 *Corrplot: correlation between each two variables.*

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## .outcome ~ sex_cd + 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(atlas_hh65plusalonepct) + s(rwjf_resident_seg_black_inx) +
##     s(rwjf_uninsured_adults_pct)
##
## Estimated degrees of freedom:
## 2.60 1.00 5.64 1.82 1.00 3.50 1.69
## 7.31 1.00 1.20  total = 36.76
##
## UBRE score: -0.02449249
```

***Figure 3**: GAM Model Selection*

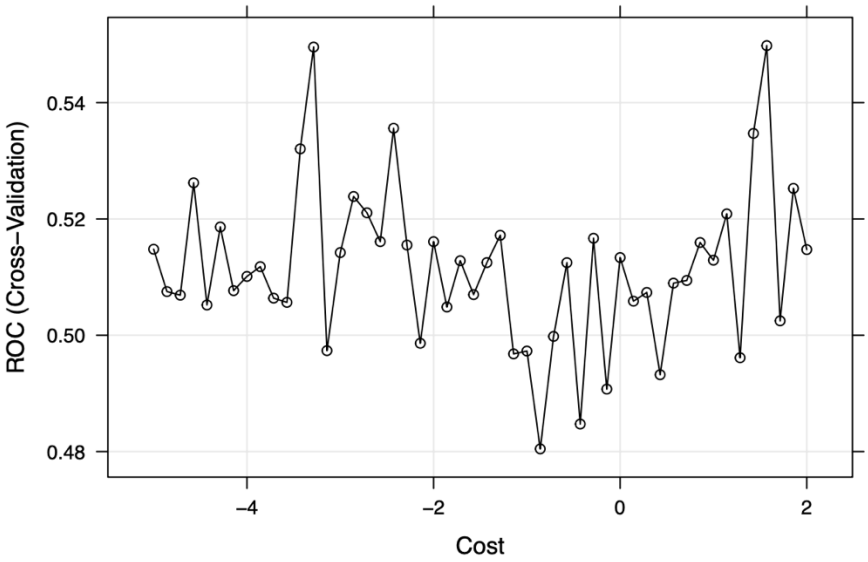

***Figure 4**: SVML Plot*

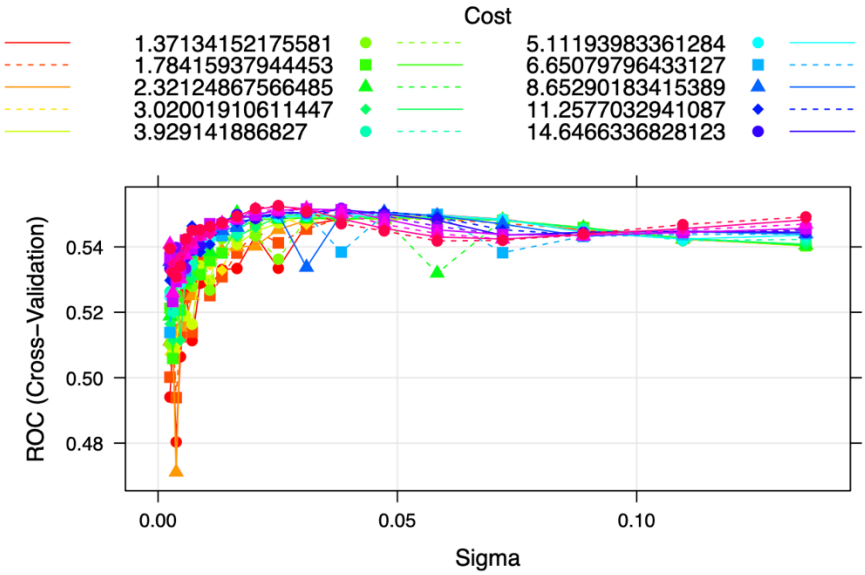

***Figure 5**: SVMR Plot*

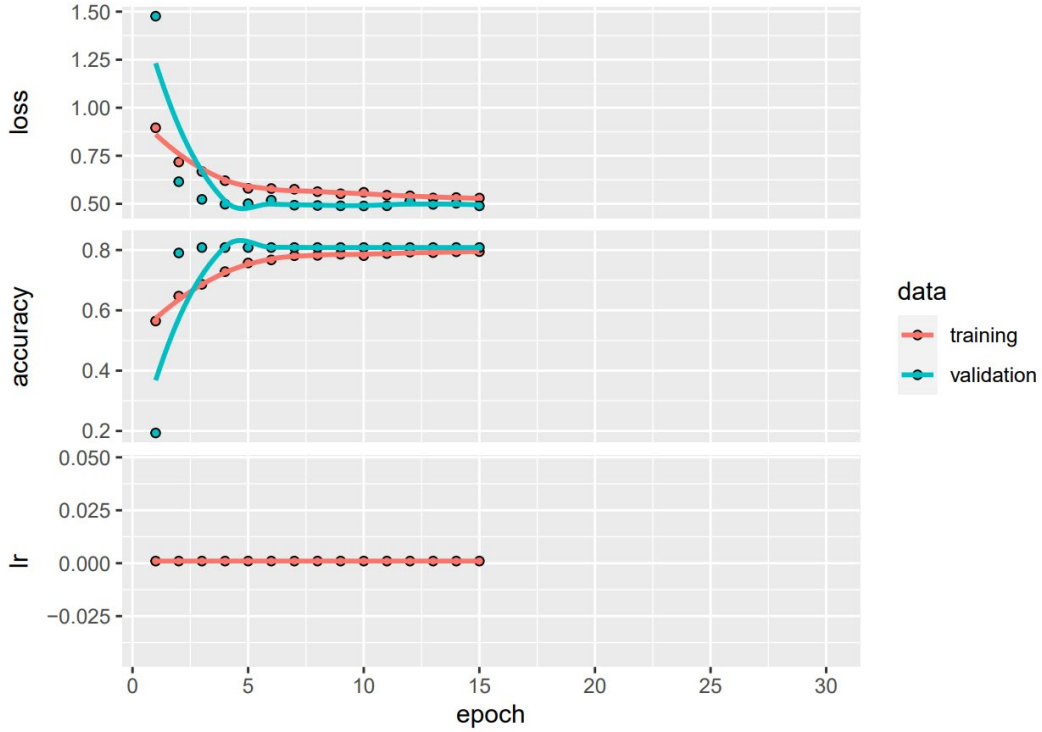

***Figure 6**: Neutral Network Training History*

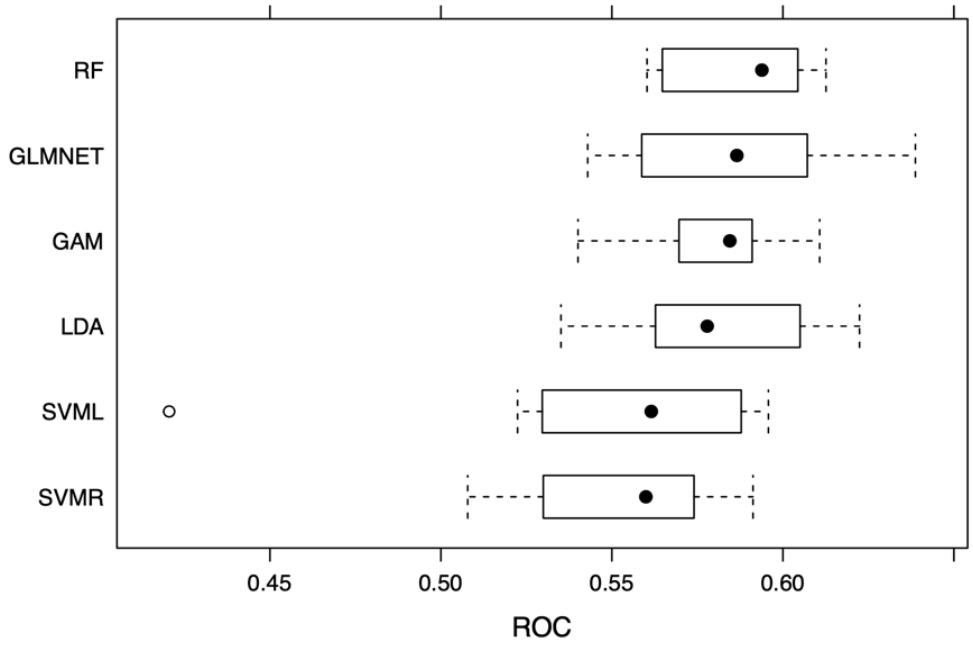

***Figure 7** Resample: ROC value of five models.*

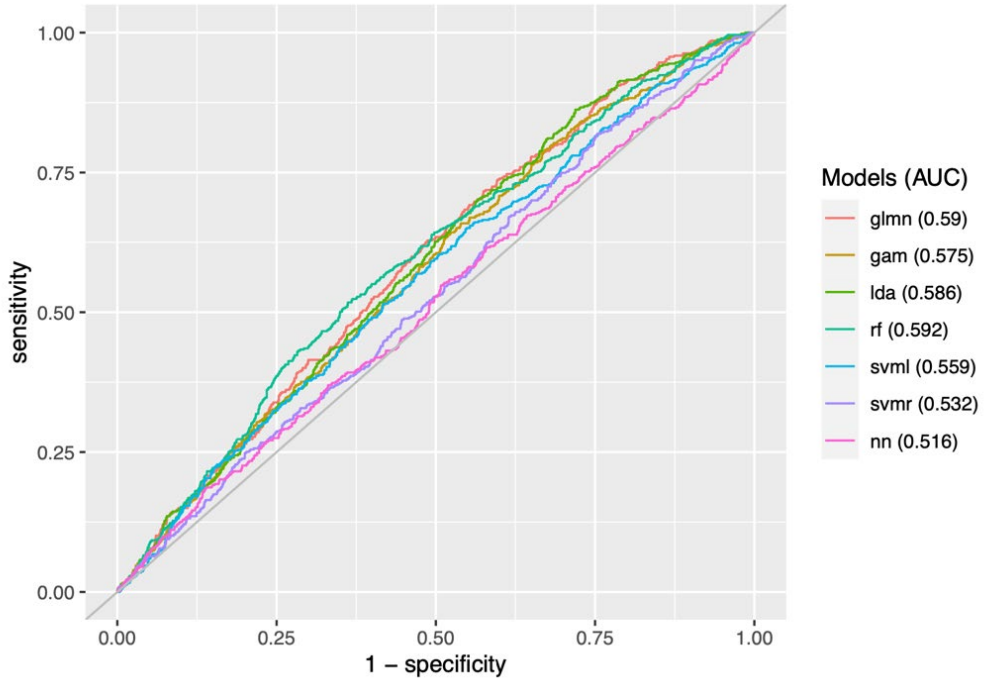

***Figure 8**: ROC Curve*

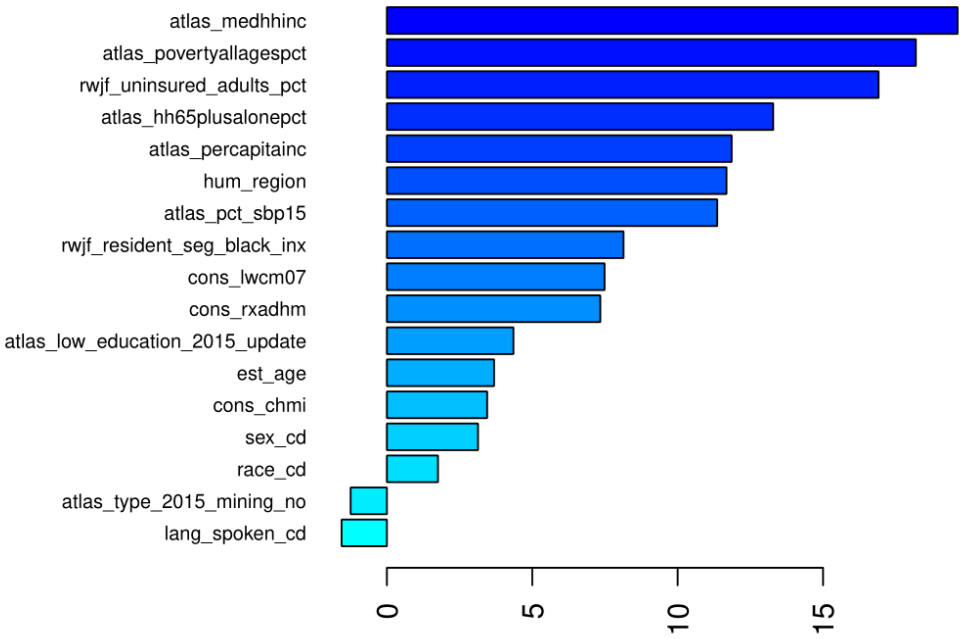

***Figure 9**: Final Model Interpretation*