Exploring the Socioeconomic Drivers of COVID-19 Vaccine Acceptance

Author: WeiJun Li (liw18@rpi.edu)

([1]Rensselaer Polytechnic Institute 110 8th St., Troy, NY, 12180 United States)

Abstract

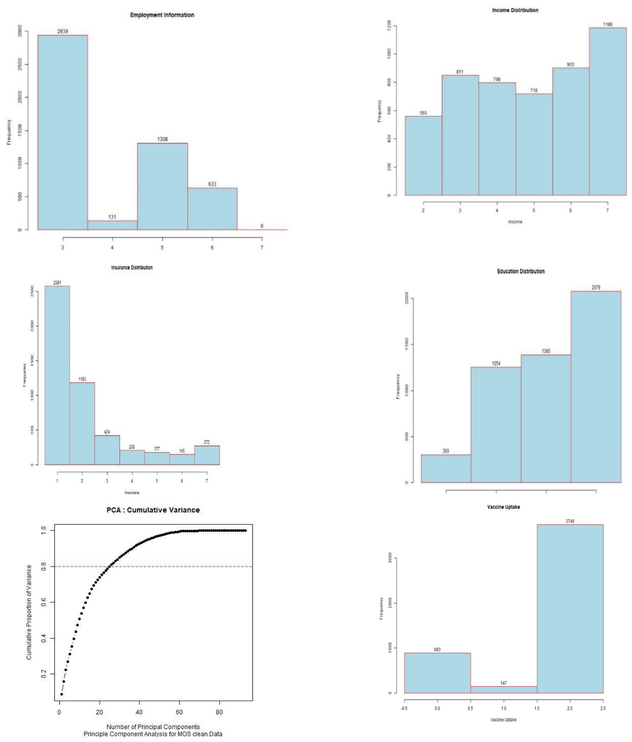
This project examines the factors that influence people's uptake of the COVID-19 vaccine and booster shots. The authors refer to the HSS COVID-19 Monthly Outcome Survey dataset found at HealthData.gov, which contains information on about 5,000 U.S. adults and covers beliefs, intentions, and behaviors related to the COVID-19 vaccine. The dataset contains information on approximately 5,000 U.S. adults covering beliefs, intentions, and behaviors related to the COVID-19 vaccine. The dataset has several features that make it suitable for analysis using classification models. The target variable is CAM5_VaccUptake, which indicates the number of times people have been vaccinated. The authors plan to identify key factors influencing vaccination intentions through exploratory data analysis (EDA) and the application of PCA, random forests, decision trees, SVM, and K-means clustering.

Motivation

The motivation of this project indicate the High vaccination rates are essential to protect susceptible populations and suggests that the government should develop policies to increase vaccination rates, such as mandatory vaccination for federal employees or workers in the health care industry.COVID-19 has caused more deaths than World War II and emphasizes the importance of vaccines and booster shots to reduce deaths.

EDA and data cleaning

For the data cleaning , remove people refuse to answer and invalid data column like data and ID for each participant. Then use the histogram to see income, education, insurance, employment factor distribution. The next is do the PCA to see how each feature contribution to my target value CAM5_VaccUptake. As we can see around 31 feature contribution over 80% of Cumulative Variance. This indicate my hypothesis is in the correct direct which some have strong effect on the target value.



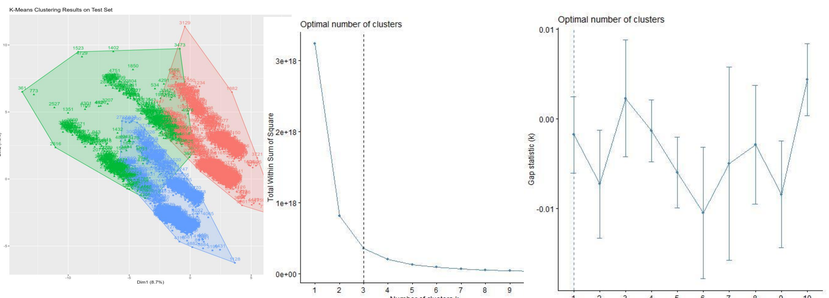
Module Development And Analysis

Import
- Necessary library class in R
- Read the dataset into R
- Clear the data

Inspect
- Explore the dataset to get descriptive results
- Find out data features and outliers

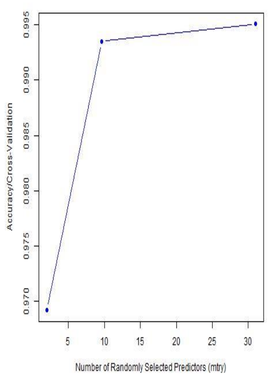
Process
- Build the model by data
- Analysis the result
- Make the conclusion for each case

K-mean



The fviz_nbclust (elbow rule) and clusGap methods were used to determine the optimal number of clusters, k. Both methods indicated that k=3 was the appropriate number of clusters. However, model performance analysis showed that there was overlap between clusters and that the sum of squares between groups was only 13% of the total sum of squares, suggesting that the model failed to effectively differentiate between different groups. Therefore, the application of k-mean clustering in this project is not satisfactory.

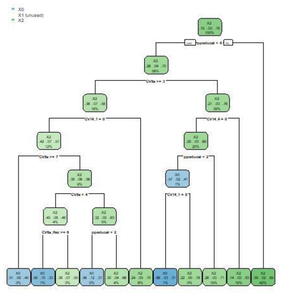
Random Forest




```
> print(confusion)
Confusion Matrix and Statistics

          Reference
Prediction  X0  X1  X2
        X0 178   1   0
        X1   0  28   0
        X2   0   0 749

Overall Statistics

               Accuracy : 0.999
                 95% CI : (0.9942, 1)
    No Information Rate : 0.7835
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.997

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: X0 Class: X1 Class: X2
Sensitivity             1.0000   0.96552    1.0000
Specificity             0.9987   1.00000    1.0000
Pos Pred Value          0.9944   1.00000    1.0000
Neg Pred Value          1.0000   0.99892    1.0000
Prevalence              0.1862   0.03033    0.7835
Detection Rate          0.1862   0.02929    0.7835
Detection Prevalence    0.1872   0.02929    0.7835
Balanced Accuracy       0.9994   0.98276    1.0000
```

This model can handle complex relationships and does not require data cleaning. By splitting 80% of the training set and 20% of the test set, as well as k-fold cross-validation, the authors determined the optimal mtry parameter to be 31, indicating that choosing 31 features for prediction is optimal. The model performs well on the test set with 99.9% accuracy and sensitivities for each group close to 1. The results also show that the accuracy is close to 99% even with only 5-10 features, indicating that only a few key features are needed for effective prediction. This suggests that vaccination rates can be significantly improved by controlling for a few key variables. The authors plan to further analyze this in depth using decision tree modeling.

Decision Tree



```
> print(confusion)
Confusion Matrix and Statistics

          Reference
Prediction  X0  X1  X2
        X0  31   2  24
        X1   0   0   0
        X2 147  27 725

Overall Statistics

               Accuracy : 0.7908
                 95% CI : (0.7636, 0.8162)
    No Information Rate : 0.7835
    P-Value [Acc > NIR] : 0.3068

                  Kappa : 0.1703

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: X0 Class: X1 Class: X2
Sensitivity             0.17416   0.00000   0.9680
Specificity             0.96658   1.00000   0.1594
Pos Pred Value          0.54386       NaN   0.8065
Neg Pred Value          0.83648   0.96967   0.5789
Prevalence              0.18619   0.03033   0.7835
Detection Rate          0.03243   0.00000   0.7584
Detection Prevalence    0.05962   0.00000   0.9404
Balanced Accuracy       0.57037   0.50000   0.5637
```

```
> print(dt_model)
CART

3833 samples
   8 predictor
   3 classes: 'X0', 'X1', 'X2'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3449, 3449, 3450, 3449, 3450, 3451, ...
Resampling results across tuning parameters:

  cp          logLoss    AUC        prAUC      Accuracy   Kappa      Mean_F1
  0.002480960  0.5953044  0.6663075  0.2581415  0.7865872  0.1789695  NaN
  0.002881120  0.5768889  0.6718998  0.2883962  0.7891962  0.1755122  NaN
  0.004201681  0.5889896  0.6333738  0.2211245  0.7837226  0.1187011  NaN
  Mean_Sensitivity  Mean_Specificity  Mean_Pos_Pred_Value  Mean_Neg_Pred_Value
  0.3801286         0.7111716         0.4288995            0.8005984
  0.3808748         0.7122070                 NaN          0.8075010
  0.3650276         0.6975727                 NaN          0.7859697
  Mean_Precision  Mean_Recall  Mean_Detection_Rate  Mean_Balanced_Accuracy
  0.4288995       0.3801286    0.2621957            0.5456501
        NaN       0.3808748    0.2638654            0.5465409
        NaN       0.3650276    0.2612489            0.5313002

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.00288112.
```

In the project, the authors used a decision tree model for classification and the best model parameters were determined after segmentation and K-fold cross-validation on 80% training set and 20% test set. The model had an accuracy of 79.08% on the test set, but performed poorly in predicting those who were vaccinated only once, with a sensitivity of 0. This may be due to the small number of people in the sample who were vaccinated only once. In addition, since the time point selected for the dataset was near the end of the epidemic, this may have affected the performance of the model, suggesting that temporal factors and changes in vaccine policy have an impact on vaccination behavior.

Support vector machine(SVM)

```
> print(confusion)
Confusion Matrix and Statistics

          Reference
Prediction  X0  X1  X2
        X0  10   2  15
        X1   0   0   0
        X2 168  27 734

Overall Statistics

               Accuracy : 0.7782
                 95% CI : (0.7505, 0.8042)
    No Information Rate : 0.7835
    P-Value [Acc > NIR] : 0.6693

                  Kappa : 0.0499

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: X0 Class: X1 Class: X2
Sensitivity             0.05618   0.00000   0.97997
Specificity             0.97815   1.00000   0.05797
Pos Pred Value          0.37037       NaN   0.79010
Neg Pred Value          0.81916   0.96967   0.44444
Prevalence              0.18619   0.03033   0.78347
Detection Rate          0.01046   0.00000   0.76778
Detection Prevalence    0.02824   0.00000   0.97176
Balanced Accuracy       0.51716   0.50000   0.51897
```

```
> print(svm_model)
Support Vector Machines with Radial Basis Function Kernel

3833 samples
   8 predictor
   3 classes: 'X0', 'X1', 'X2'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3449, 3449, 3450, 3449, 3450, 3451, ...
Resampling results across tuning parameters:

  C     logLoss    AUC        prAUC      Accuracy   Kappa      Mean_F1
  0.25  0.5872341  0.6407227  0.4203355  0.7873685  0.1034306  NaN
  0.50  0.5872112  0.6489784  0.4199666  0.7855428  0.0933852  NaN
  1.00  0.5850874  0.6455521  0.4229884  0.7871088  0.1013211  NaN
  Mean_Sensitivity  Mean_Specificity  Mean_Pos_Pred_Value  Mean_Neg_Pred_Value
  0.3601912         0.6911886                 NaN          0.8111899
  0.3576411         0.6886990                 NaN          0.8057759
  0.3593568         0.6904782                 NaN          0.8089055
  Mean_Precision  Mean_Recall  Mean_Detection_Rate  Mean_Balanced_Accuracy
  NaN             0.3601912    0.2624562            0.5256899
  NaN             0.3576411    0.2618476            0.5231700
  NaN             0.3593568    0.2623696            0.5249175

Tuning parameter 'sigma' was held constant at a value of 0.1336754
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.1336754 and C = 0.25.
```

The dataset was first divided into 80% training set and 20% test set and the optimal C and sigma parameters were determined by K-fold cross validation. Eight features were selected to train the SVM model using Gaussian kernel. The test results show an accuracy of 78% when sigma is 0.1336754 and C is 0.25. Application to the test set showed that the model was inaccurate in predicting the category of only one vaccination with an accuracy of 77.82% and a 95% confidence interval ranging from 0.7505 to 0.8042.The p-value of McNemar's Test was very low indicating significant differences in prediction accuracy between categories. The SVM model was biased in predicting some of the categories, which could be due to training imbalance in the categories of the data or improper selection of model parameters.

Summary

Factors Influencing COVID-19 Vaccination Behavior A variety of classification models were used for this, including random forests, decision trees, support vector machines (SVMs), and K-means clustering, and factors such as income, education, insurance, employment, and good behavior were found to have a significant effect on vaccination. Adjustments in data preprocessing and model selection during the project had a significant impact on the final results. In future studies, there are plans to explore additional data balancing techniques and the application of more advanced models, such as deep learning networks, as well as the implementation of more comprehensive feature engineering. Despite the positive preliminary results, more research and analysis are needed to more fully understand the factors that influence vaccination behavior.

Resources:

[1] U.S. Department of Health & Human Services.. HHS COVID-19 Monthly Outcome Survey Wave 28. HealthData.gov. Retrieved from https://healthdata.gov/Health/HHS-COVID-19-Monthly-Outcome-Survey-Wave-09/6itx-ccwh/about_data

[2] U.S. Census Bureau. (2022, March 17). Educational Attainment in the United States: 2021. Retrieved from https://www.census.gov/newsroom/press-releases/2022/educational-attainment.html

[3]Peter G. Peterson Foundation. (2023, November). The Share of Americans Without Health Insurance in 2022 Matched a Record Low. Retrieved from https://www.pgpf.org/blog/2023/11/the-share-of-americans-without-health-insurance-in-2022-matched-a-record-low

[4]Centers for Disease Control and Prevention. (n.d.). Weekly Updates by Select Demographic and Geographic Characteristics. Retrieved from https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm

Link to GitHub repository:
https://github.com/wei-jun7/DataAnalytics2023_WeijunLi