GPH-GU 2338: Final project report

Application of Machine Learning Algorithms to Identify People with Economic Vulnerability

Sooyoung Kim

Jessica Randazzo

Introduction

Since the World Health Organization (WHO) declared COVID-19 as a Public Health Emergency of International Concern (PHEIC) on January 30, 2020 1 until now, the COVID-19 pandemic continues to cause significant global disruption across sectors ranging from healthcare to education to the economy.[ref] Several studies across the globe suggested that the burden of the pandemic is disproportionately affecting the vulnerable population both in economic and health perspective 2-5.

In the United States, mounting evidence suggests that vulnerable populations became even more vulnerable during the pandemic, including the racial minorities, those who are uninsured, and/or in poverty. 6-9 This potentially contributes to the widening health- and socioeconomic disparities in the country. 6,10,11

The most evident adverse impact of the pandemic, aside from health outcomes, was observed in the economic aspect.[ref] The economic shock of the pandemic not only affects businesses and industries, but also directly affects individuals and households.[ref] Studies published since the beginning of the COVID-19 epidemic have consistently demonstrated the issue of housing-, job-, and food insecurity, among many other forms of financial hardships.[ref]

Related Work

Addressing socioeconomic disparity is important in addressing health disparities, as social determinants are associated with various morbidity and mortality outcomes.[ref] While the federal tax relief, so-called Stimulus Check, a means-tested benefit program, has been deployed to release the economic burden of individuals and households across the US, it is unclear whether this benefit can help narrow the disparity. Previous studies suggested that a means-tested benefit was not associated with improving the health outcomes, while some entitlement-based economic benefits were.[ref] In addition, even if such benefits exist, identifying the most vulnerable population and quantifying the economic burden or loss they’re suffering in a timely manner is also challenging.

In order to better address the disparate economic vulnerability within the population, we aim to identify the vulnerable sub-populations that are significantly affected by the COVID-19 pandemic using the close-to-real-time data of national panel surveys. The aim of this study is to support better allocation of the resources to aid the economic recovery in manners that reduce the socioeconomic disparity.

Methods

Outcome variable

Four variables were identified to be related to economic vulnerability. The four variables are financial hardship in the past 7 days, housing insecurity, food insecurity, and use of stimulus check on essential expenses. This study focused on the outcome “financial hardship” which was assessed based on the binary response to the question ‘In the last 7 days, how difficult has it been for your household to pay for usual household expenses, including but not limited to food, rent or mortgage, car payments, medical expenses, student loans, and so on?`.

Predictors

16 predictors representing respondent’s demographic and socioeconomic characteristics are included. Age at the time of survey was included as a continuous variable. Other demographic characteristics included gender, race and ethnicity, marital status, urban/rurality of the residence, region of the residence. Socioeconomic characteristics included the highest educational attainment, household income, type of health insurance owned, employment type, and the type of housing. We also included the recent history of job loss, COVID-19 infection history, whether or not the respondent received the unemployment and social security insurance benefit during the pandemic, as well as the summed score of 4 patient health questionnaire (PHQ-4) measuring respondents’ mental health status.17

Fitting the Machine Learning Algorithm

In order to find the best ML algorithm for classification, we fit the data using 5 different methods – logistic regression, support vector machine, linear discriminant analysis, quadratic discriminant analysis, and decision trees, which also includes bagging and random forest. Based on the best model derived from each method, we compared their test error rate (misclassification rate on the test data) and chose the final algorithm and the model.

Logistic regression: For each outcome variable, logistic regression was fit using the 16 predictors. Then we used the AIC-based subset selection method to choose the 10 best models. 10-fold cross-validation was used to calculate the cross-validation error, which will be the criteria to choose the final model. For classification of the outcome, cut-off value of 0.5 was used throughout the entire process.

Support vector machine (SVM): Due to the limited computational power, SVM was fitted using the subset of the training set (N = 3000, corresponding to roughly 10% of the training set). Using the linear, radial and polynomial kernel, ranges of values on cost (0.1, 1, 10, 100, 1000) and the gamma (0.5, 1, 2, 3, 4), for the radial and polynomial kernel only, were tested. Best model was selected using the 10-fold cross validation and the minimum misclassification rate.

Linear Discriminant Analysis (LDA): We did an LDA on our data to compare its performance with other popular methods. Because our data does not characteristically fit the reasons for choosing an LDA, results were interpreted carefully. We predicted financial hardship using all of the predictors on the training dataset. 89% of the training observations corresponded to no financial hardship in the past 7 days.

Quadratic Discriminant Analysis (QDA): We also tried QDA for our data because it is more flexible than LDA, assumes each class has its own covariance matrix, and our training set is large enough that we don’t have to worry about the variance of the classifier.

Decision Tree / Bagging / Random Forest: Lastly, we used classification trees, bagging, and random forests on our data. Many nodes result in a prediction of no financial hardship in the last 7 days. Cross-validation was conducted to see if the tree could be pruned further and pruning was done accordingly. For the bagging method, all variables in the training data set were used at each split, and we tested its performance on the test data by calculating the misclassification rate. In our random forest model, 4 variables were tried at each split. We also obtained the variable importance plot.

Data and Experiment Setup

US Census bureau’s household pulse survey collects responses from nationally representable samples on various COVID-19-related questions, including the impact of pandemic on individuals’ and households’ health, economic, and education, among many others.12 Details of the technical documentation is provided elsewhere.13 To date, many studies adopted the subset of the household pulse survey to generate close-to-real-time evidence on how the current pandemic is affecting the nation’s households.14-16

For this study, we have limited the scope of data to the 25th wave of the survey, which was conducted between February 17 – March 1, 2021. For all subsequently described analyses, we used the division of training and test sets using the 75:25 ratio. For preprocessing, we checked which predictor variables have missing observations. Among those with missing observations, we checked whether they are missing at random (MAR) or missing completely at random (MCAR) by using the Pearson correlation matrix for continuous variables and the logistic regression for the categorical variables. For this analysis, we did not conduct further imputation to handle the predictors MAR and decided to limit our analysis to the subset without any missing observations.

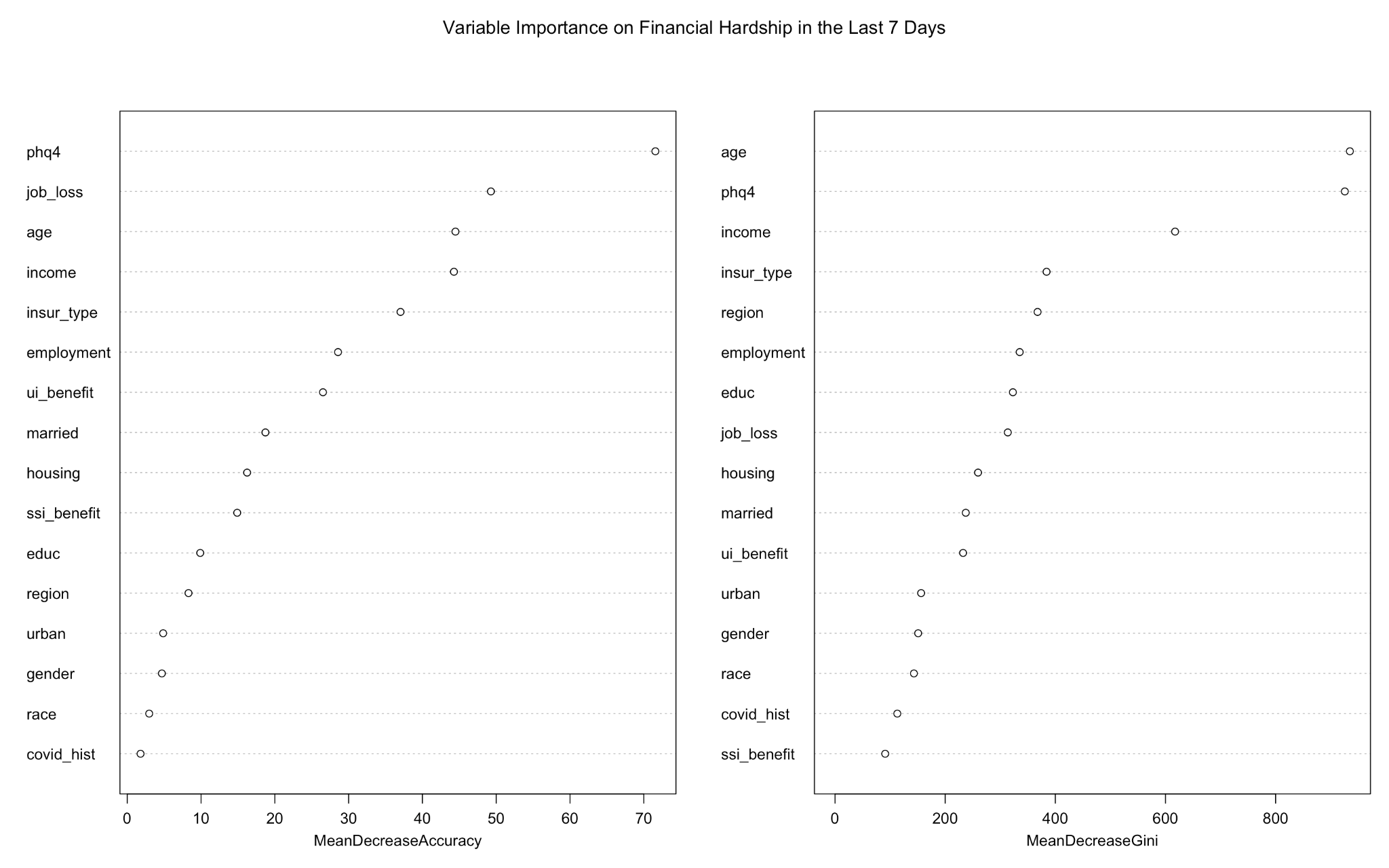
Results

Missingness diagnostics

All predictors contained some missing observations. Among them, age, gender, race, educational attainment, urban/rurality of residence, PHQ4 score, and the region of residence contain observations that are missing completely at random (MCAR). Household income, marital status, type of employment, receiving of unemployment and social security benefit, type of insurance owned, type of housing, recent COVID-19 infection history, and job loss during the pandemic seem to be missing at random (MAR). Detailed results of the missingness diagnostics are summarized in the supplementary material (Table S1 – S39). As mentioned previously, we did not conduct further imputation to handle the predictors MAR, and used the subset of the data without any missing observations for all predictors (N=40,460).

ML algorithm performance

Logistic regression model resulted in the least classification error rate (9.4%) amongst all ML algorithms. It also resulted in the smallest false omission rate (8.0%), which is important in order not to miss any vulnerable individuals in the scope of potential interventions. Support vector machine only converged for the radial kernel, however, the machine classified all observations to have no financial hardship. On our LDA model, there was a 10.3% test error rate where the QDA model showed a 18.7% test error rate. Either model is not the best for our data considering the structure of our data and the assumptions needed for these methods. The classification trees showed an error rate of 10.9% with 7 leaves including the variables job loss, income, phq4, and insurance type. Results showed that a tree with 7 or 4 nodes produced the smallest cross-validation errors. Pruning the tree to 4 nodes showed similar error rates with variables job loss, phq4, and insurance type. For simplicity, we kept the 4-node tree for further comparison with other algorithms. The misclassification rate for bagging was 10.4% which was slightly lower than the error rate obtained from the classification tree. Random forests showed a misclassification rate of 10%, leading to the smallest error rate out of the tree based methods. Additionally, this method resulted in phq4, job loss, age, income, insurance type, and employment as the most important predictors.



Classification performance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ML algorithm | Method used for validation | Test error rate (misclassification rate) | False discovery rate (1-PPV) | False omission rate (1-NPV) |
| Logistic regression | 10-fold cross validation | 9.4% | 35% | 8.0% |
| LDA | Prediction on the test error | 10.3% |  |  |
| QDA | ^ | 18.7% |  |  |
| SVM - radial | 10-fold cross validation and comparison of different kernels | 11.0% | Unable to calculate | 11.4% |
| SVM - linear | ^ |  |  |  |
| Trees |  |  |  |  |
| Classification Trees | 10-fold cross validation and pruning | 10.7% |  |  |
| Bagging | Prediction on test error | 10.4% |  |  |
| Random Forests | ^ | 10% |  |  |

Discussion

Our analysis using random forests showed X variables were significant in predicting financial hardship due to the pandemic.

We decided to focus on financial hardship as the sole outcome of interest, neglecting three other potential outcome variables (housing insecurity, food insecurity, and the use of stimulus check on essential household items). Repeating the same analysis to these three outcomes in the future will provide further insight to understand the different aspects of economic vulnerability and its distribution across the population.

We also neglected LASSO and Ridge regression methods to be part of our study mainly due to the limitation of computational power. We fully acknowledge the possibility that model selection using LASSO, including all variables included in the original household pulse survey dataset, may result in a model with better prediction performance.

Contributions

Sooyoung Kim conducted data cleaning, preliminary inspection, and the recoding, as well as the machine learning algorithm fitting using the logistic regression and the support vector machine method. Jessica Randazzo was responsible for the linear/quadratic discriminant analysis and the decision trees (classification tree, bagging, random forest). Both equally contributed to the literature review, synthesis of findings, writing of the report, and the preparation of the presentation. The R project to conduct all the analyses described in this report is available in the github repository (<https://github.com/sooyoung1021/GPH2338_project>)

References

1. World Health Organization. COVID 19 Public Health Emergency of International Concern (PHEIC)

Global research and innovation forum: towards a research roadmap. 2020. <https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-(pheic)-global-research-and-innovation-forum>. Published 12 February 2020. Accessed 19 October 2020.

2. Grooms; J, Ortega; A, Rubalcaba JA-A. The COVID-19 public health and economic crises leave vulnerable populations exposed. In. *Brookings*2020.

3. Qian Y, Fan W. Who loses income during the COVID-19 outbreak? Evidence from China. *Research in Social Stratification and Mobility.* 2020;68.

4. Mottaleb KA, Mainuddin M, Sonobe T. COVID-19 induced economic loss and ensuring food security for vulnerable groups: Policy implications from Bangladesh. *PLoS One.* 2020;15(10):e0240709.

5. Ataguba JE. COVID-19 Pandemic, a War to be Won: Understanding its Economic Implications for Africa. *Appl Health Econ Health Policy.* 2020;18(3):325-328.

6. McLaren J. *Racial Disparity in COVID-19 Deaths: Seeking Economic Roots with Census data.*: US National Bureau of Economic Research; June 2020 2020.

7. Wolfson JA, Leung CW. Food Insecurity and COVID-19: Disparities in Early Effects for US Adults. *Nutrients.* 2020;12(6).

8. Woolhandler S, Himmelstein DU. Intersecting U.S. Epidemics: COVID-19 and Lack of Health Insurance. *Ann Intern Med.* 2020;173(1):63-64.

9. Webb Hooper M, Napoles AM, Perez-Stable EJ. COVID-19 and Racial/Ethnic Disparities. *JAMA.* 2020;323(24):2466-2467.

10. Azar KMJ, Shen Z, Romanelli RJ, et al. Disparities In Outcomes Among COVID-19 Patients In A Large Health Care System In California. *Health Aff (Millwood).* 2020;39(7):1253-1262.

11. Wiemers EE, Abrahams S, AlFakhri M, Hotz VJ, Schoeni RF, Seltzer JA. Disparities in Vulnerability to Severe Complications from COVID-19 in the United States. *medRxiv.* 2020.

12. US Census Bureau. Measuring Household Experiences during the Coronavirus Pandemic. In: US Census Bureau, ed2021.

13. JF F, J H-C, A T, et al. *Design and Operation of the 2020 Household Pulse Survey.* U.S. Census Bureau;2020.

14. Berkowitz SA, Basu S. Unemployment Insurance, Health-Related Social Needs, Health Care Access, and Mental Health During the COVID-19 Pandemic. *JAMA Intern Med.* 2020.

15. Schanzenbach; D, Pitts A. *How Much Has Food Insecurity Risen? Evidence from the Census Household Pulse Survey.* Northwestern Institute for Policy Research; June 10, 2020 2020.

16. Cutler DM, Summers LH. The COVID-19 Pandemic and the $16 Trillion Virus. *JAMA.* 2020;324(15):1495-1496.

17. Lowe B, Wahl I, Rose M, et al. A 4-item measure of depression and anxiety: validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *J Affect Disord.* 2010;122(1-2):86-95.