Cerner Enviza
an Oracle company

Unsupervised clustering and binary classification analysis for predicting COVID syndrome (Long COVID) in US adults

Kyla S. Finlayson

Vicky W. Li

Oracle Life Sciences, Real World Evidence, New York, United States of America

May 8, 2023

Contents

1
Introduction

Background

Condition overview

Post-COVID syndrome, or "long COVID" is an emerging chronic condition in the United States that will likely have a substantial impact on the healthcare system over the next decade.

- Preliminary estimates from the annual Household Pulse Survey reveal that "nearly **1 in 5** American adults who have had COVID-19 still have long COVID", making up **~13% of the total US adult population**[1]

Symptoms

Long COVID is "a syndrome characterized by persistent symptoms and/or delayed or long-term complications beyond 4 weeks from the onset of (COVID-19) symptoms."[2]

- People with long COVID can experience a **wide range of symptoms** including respiratory, heart, neurological, and digestive ailments.[3]
- Additionally, it can elicit many mental health issues including anxiety, depression, mood changes, brain fog, and sleep disturbances.[3]

Study objectives

USING RECENT PATIENT-REPORTED DATA IN THE U.S. TO:

1.

Explore whether unique **symptom clusters** exist among patients who have reported experiencing Long COVID

2.

Predict Long COVID based on demographic profile, acute COVID-19 symptoms experienced, comorbidities, and health characteristics

- Identify the best model at predicting Long COVID
- Identify the **top risk factors** for Long COVID

2
Methods

Data

Sample

2022 National Health and Wellness Survey

- Internet-based survey administered to a general population panel
- Adult participants age 18+
- Representative of US Census on age, gender, and ethnicity
- 75,261 participants

Inclusion Criteria

Objective 1: Explore whether unique symptom clusters may exist among respondents who have reported experiencing Long COVID

- Reported ever experiencing Long COVID: **N = 2,665**

Objective 2: Predict Long COVID and identify the top risk factors

- Reported ever experiencing COVID-19: **N = 13,953**

Variables

Outcome: Long COVID

- Binary variable
- Defined as whether a respondent was ever told by a doctor that they have Long COVID

Predictors (202 in total)

- Socio-demographic characteristics (6)
- Health-related characteristics (8)
  - Body mass index, smoking behavior, and more
- COVID-19 symptoms (27)
- Other COVID-19 specific characteristics (10)
  - COVID-19 related hospital visits, vaccination status, and more
- Comorbidities (151)
  - Including respiratory conditions, heart conditions, mental health conditions, diabetes, cancer, and more

Objective 1: Clustering analysis methods

Features	**27 binary COVID-19 symptom variables**
Clustering method	**K-Medoids clustering**
Distance measure	**Jaccard dissimilarity**
Dimension reduction	**t-SNE dimension reduction**
Selection criteria for number of clusters	**Average silhouette width**

Objective 2: Binary classification models

LASSO logistic regression, Random Forest classification, XGBoost classification

All three models:

- Selection criterion: AUC
- Training data: 80%, test data: 20%
- 10-fold cross-validation to find optimal hyperparameters

Performance measures:

- Performance measures: AUC, accuracy, sensitivity, and specificity

LASSO Logistic Regression

- R glmnet package[5]
- Regularization parameter λ = 0.004009799
- Exponentiated standardized beta estimates for feature importance

Random Forest Classification

- R ranger package[6]
- Hyperparameters:
  - Mtry = 10, min.node.size = 5, num.trees = 300
- Mean decrease in node impurity for feature importance

XGBoost Classification

- R xgboost package[7]
- Hyperparameters:
  - Max.depth = 6, eta = 0.05, colsample_bytree = 0.25, min_child_weight = 25, lambda = 0.8, nrounds = 200
- SHAP values for feature importance

3
Results

Objective 1: Cluster analysis results for COVID-19 symptoms

	Cluster 1 (N = 282)	Cluster 2 (N = 423)	Cluster 3 (N = 393)	Cluster 4 (N = 384)	Cluster 5 (N = 242)	Cluster 6 (N = 341)	Cluster 7 (N = 309)	Cluster 8 (N = 148)	Cluster 9 (N = 143)
Fatigue / tired	16.0%	**84.2%**	**63.4%**	24.2%	**85.1%**	15.5%	14.2%	0.0%	0.0%
Cough	0.0%	**83.2%**	**99.5%**	**100.0%**	**66.5%**	0.0%	0.0%	**100.0%**	0.0%
Difficulty breathing or shortness of breath	**98.6%**	**83.2%**	44.3%	22.4%	**57.9%**	0.6%	0.0%	0.0%	0.0%
Fever or chills	11.0%	**83.0%**	**99.8%**	6.0%	**60.7%**	**56.6%**	0.0%	0.0%	0.0%
Muscle or body aches	9.6%	**82.3%**	**59.3%**	19.8%	**67.8%**	9.7%	15.9%	0.0%	0.0%
Headache	9.6%	**80.9%**	**64.1%**	25.0%	**65.7%**	14.4%	19.4%	0.0%	0.0%
New confusion (brain fog)	16.3%	**71.9%**	20.6%	13.8%	49.6%	8.2%	23.0%	0.0%	0.0%
Congestion or runny nose	8.5%	**71.2%**	38.2%	11.7%	45.9%	8.2%	7.1%	0.0%	0.0%
Sore throat	9.9%	**69.3%**	44.0%	14.6%	37.6%	7.3%	14.9%	0.0%	0.0%
General pain / discomfort	1.8%	**67.4%**	20.9%	5.0%	45.9%	1.5%	3.2%	0.0%	0.0%
New loss of smell or taste	14.9%	**66.2%**	**50.9%**	21.1%	44.6%	46.9%	3.9%	0.0%	0.0%
Dizziness / lightheadedness on standing	5.3%	**66.0%**	15.5%	7.0%	41.7%	3.8%	3.9%	0.0%	0.0%
Sleep problems	7.1%	**53.2%**	17.6%	11.2%	37.6%	6.7%	6.5%	0.0%	0.0%
Pale, gray, or blue-colored skin or nail beds	11.0%	15.8%	3.3%	18.2%	10.3%	13.5%	10.0%	0.0%	**100.0%**

*Thirteen other COVID-19 symptoms were used in the clustering analysis, but are not included in the table because they were experienced less among all clusters

Objective 1: Cluster analysis results for COVID-19 symptoms

Demographic characteristics and comorbidities across clusters

	Cluster 1 (N = 282)		Cluster 2 (N = 423)		Cluster 3 (N = 393)		Cluster 4 (N = 384)		Cluster 5 (N = 242)		Cluster 6 (N = 341)		Cluster 7 (N = 309)		Cluster 8 (N = 148)		Cluster 9 (N = 143)	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Age (Mean ± SD)	38.26 ± 11.65		**47.08 ± 15.34**		46.55 ± 16.54		42.25 ± 13.90		49.19 ± 16.10		38.46 ± 11.71		37.20 ± 10.15		37.93 ± 9.31		39.27 ± 7.73	
Gender: Female	94	33.3%	**330**	**78.0%**	238	60.6%	143	37.2%	174	71.9%	118	34.6%	119	38.5%	38	25.7%	54	37.8%
Annual Income: <$75,000	64	22.7%	**253**	**59.8%**	196	49.9%	110	28.7%	147	60.7%	61	17.9%	66	21.4%	30	20.3%	33	23.1%
BMI: 35+	18	6.4%	**107**	**25.3%**	79	20.1%	36	9.4%	60	24.8%	21	6.2%	13	4.2%	4	2.7%	20	14.0%
COVID-19 vaccination	265	94.0%	**310**	**73.3%**	312	79.4%	354	92.2%	199	82.2%	307	90.0%	280	90.6%	146	98.7%	140	97.9%
Allergies	52	18.4%	**250**	**59.1%**	175	44.5%	102	26.6%	122	50.4%	53	15.5%	42	13.6%	10	6.8%	3	2.1%
Asthma	26	9.2%	**122**	**28.8%**	56	14.3%	47	12.2%	36	14.9%	28	8.2%	26	8.4%	8	5.4%	6	4.2%
Anxiety	22	7.8%	**205**	**48.5%**	109	27.7%	51	13.3%	106	43.8%	26	7.6%	17	5.5%	5	3.4%	2	1.4%
Chronic bronchitis	5	1.8%	**53**	**12.5%**	25	6.4%	19	5.0%	13	5.4%	13	3.8%	12	3.9%	2	1.4%	3	2.1%
Depression	18	6.4%	**204**	**48.2%**	116	29.5%	57	14.8%	94	38.8%	23	6.7%	25	8.1%	4	2.7%	3	2.1%
Migraine	28	9.9%	**144**	**34.0%**	84	21.4%	48	12.5%	56	23.1%	28	8.2%	19	6.2%	5	3.4%	3	2.1%

Objective 2: Binary classification results for predicting long COVID

Model performance

Model	AUC	Accuracy	Sensitivity	Specificity
LASSO Logistic Regression	0.863	0.847	0.699	0.881
Random Forest Classification	0.871	0.847	0.724	0.875
XGBoost Classification	**0.878**	**0.857**	**0.716**	**0.889**

Objective 2: Binary classification results for predicting long COVID

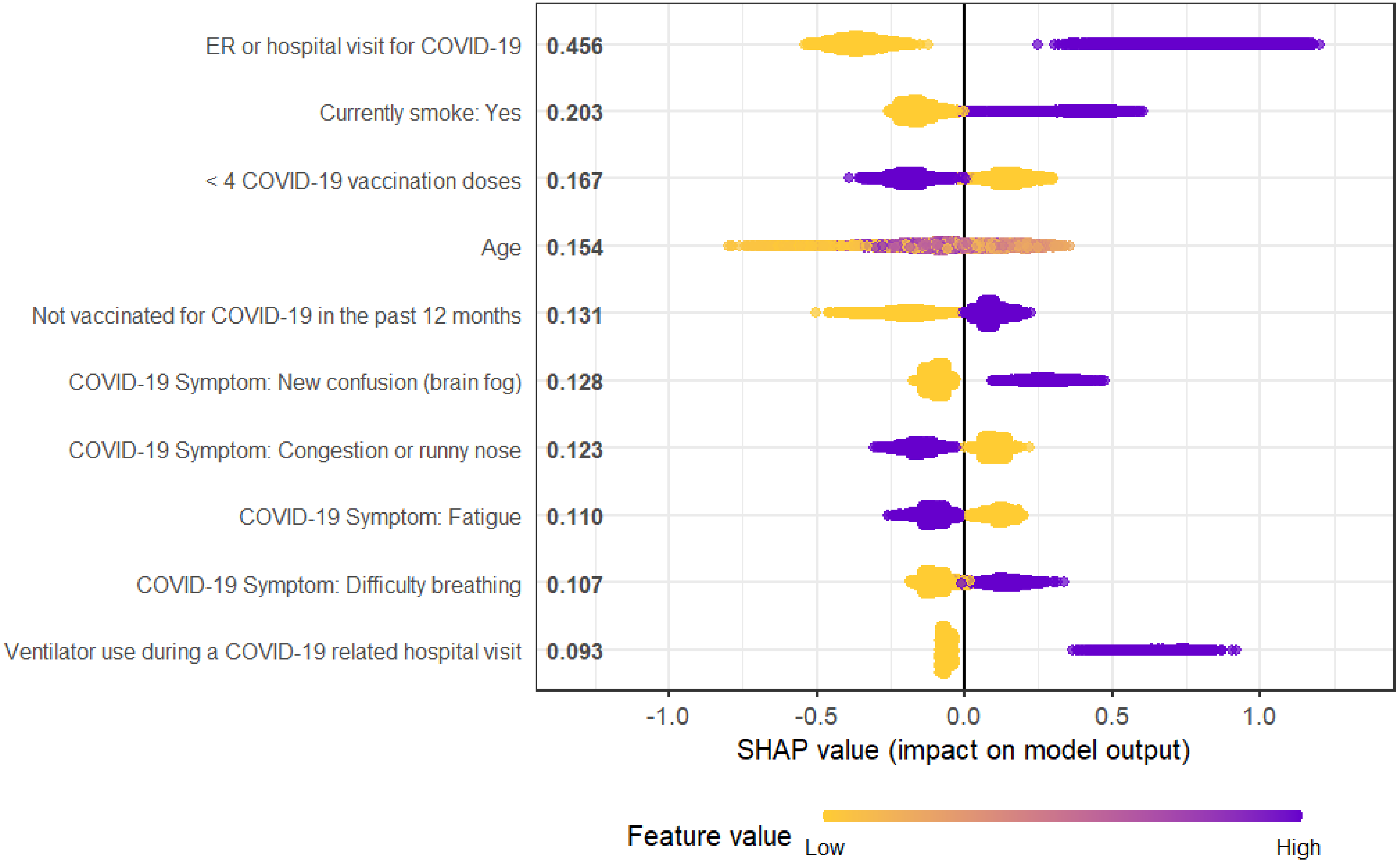
Feature importance (top 10) of binary classification models

Selected as a top feature in 3 or more models

Rank	LASSO	Random Forest	XGBoost
1	**Experienced a COVID-19 related hospital or ER visit**	**Experienced a COVID-19 related hospital or ER visit**	**Experienced a COVID-19 related hospital or ER visit**
2	**Currently Smoke: Yes**	**Admitted to the ICU during a COVID-19 related hospital visit**	**Currently Smoke: Yes**
3	**Diagnosed with Schizophrenia**	**Currently Smoke: Yes**	**< 4 COVID-19 vaccination doses**
4	COVID-19 Symptom: Memory loss	Age	Age
5	Was put on a ventilator during a COVID-19 related hospital visit	Was put on a ventilator during a COVID-19 related hospital visit	Not vaccinated for COVID-19 within the past 12 months
6	COVID-19 Severity: Severe	Number of days exercised within the past month	COVID-19 Symptom: New confusion (brain fog)
7	COVID-19 Symptom: New confusion (brain fog)	Was put on oxygen support during a COVID-19 related hospital visit	COVID-19 Symptom: Congestion or runny nose
8	Most recent blood pressure: Low	Charlson Comorbidity Index	COVID-19 Symptom: Fatigue
9	Diagnosed with Fibromyalgia	Not vaccinated for COVID-19 within the past 12 months	COVID-19 Symptom: Difficulty breathing or shortness of breath
10	Diagnosed with Idiopathic Hypersomnia	Ever Smoked: Yes	Was put on a ventilator during a COVID-19 related hospital visit

Objective 2: Binary classification results for predicting long COVID

Feature importance (top 10) of XGBoost model



4
Conclusion

Conclusion

The cluster analysis identified a unique subgroup of patients that may be **more susceptive** to developing multiple symptoms and experiencing more severe long COVID.

In predicting Long COVID, **XGBoost** performed slightly better than random forest and LASSO in terms of AUC and accuracy

Different predictors were selected:

- LASSO logistic regression: more **comorbidity** variables
- Random Forest + XGBoost: more **health behavior and COVID-19** related variables

Conclusion

Limitations

- Cross-sectional data, no causal inferences can be made
- Self-reported diagnoses are not verified through clinician or medical claims
- COVID-19 cases may be underreported
  - Only 18.5% of total respondents reported ever experiencing COVID-19, while seroprevalence data from the CDC suggests that 63.7% of US adults have had COVID-19 at some point[8]
- Model performance could be improved with data that is more balanced towards COVID-19 patients

Future Improvements

- **Features:** Including more features specific to Long COVID rather than acute COVID-19
- **Timing:** Recent health experience would be more desirable than conditions diagnosed in more distant past

References

1. Centers for Disease Control and Prevention. "Nearly one in five American adults who have had COVID-19 still have “long COVID.”." Press release (2022).
2. Nalbandian, Ani, Kartik Sehgal, Aakriti Gupta, Mahesh V. Madhavan, Claire McGroder, Jacob S. Stevens, Joshua R. Cook et al. "Post-acute COVID-19 syndrome." Nature medicine 27, no. 4 (2021): 601-615.
3. Centers for Disease Control and Prevention. "Long Covid or post-covid conditions." (December 2022) https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html
4. R Core Team (2022). R (Version 4.2.2): A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
5. Friedman J, Tibshirani R, Hastie T (2010). “Regularization Paths for Generalized Linear Models via Coordinate Descent.” Journal of Statistical Software, 33(1), 1–22. doi:10.18637/jss.v033.i01.
6. Wright MN, Ziegler A (2017). “ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.” Journal of Statistical Software, 77(1), 1–17. doi:10.18637/jss.v077.i01.
7. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785
8. Clarke KE, Jones JM, Deng Y, et al. Seroprevalence of Infection-Induced SARS-CoV-2 Antibodies — United States, September 2021–February 2022. MMWR Morb Mortal Wkly Rep 2022;71:606-608. DOI: http://dx.doi.org/10.15585/mmwr.mm7117e3external icon.

Cerner Enviza
an Oracle company

Thank you

Kyla S. Finlayson
Kyla.Finlayson@Oracle.com

Vicky W. Li
Vicky.L.Li@Oracle.com