Supervised Machine Learning to Identify Social Economic Behavioral Healthcare Risks for COVID-19 Related Mortality and Inform Treatment and Prevention

Brian Griner[1] Matthew Ye[2*], Chelsea Jin[3]

[1]Learning Labs, Data Science & Learning Systems LLC, grinerpb@protonmail.com [2]University of California at Berkeley, my1@berkeley.edu [3]Bristol Myers Squibb, chelsea.jin@bms.com

Introduction

- Machine Learning (ML): an artificial intelligence technique that can be used to design and train software algorithms to learn from and act on data[1].
- Supervised Learning: in ML, a class of systems and algorithms that determine a predictive model using data points with known outputs[2].
- COVID-19: a severe public health event that impacts globally
- Up to April 13, 2020, the total number of the confirmed cases was 576,774, and the mortality rate was 4.05%, national-wise[3].
- Up to Aug 27, 2020, the national mortality rate still remained 3.09%, and the aggregated number of the confirmed infections was 5.80 million, in accordance with CDC guidelines as of April 14[3].

Objectives

- To use supervised learning algorithms to identify key social-economic behavioral healthcare risks that may impact COVID-19 mortality rate and mortality rate change (increase vs. decrease).

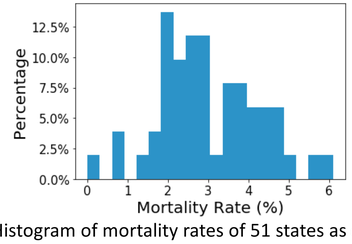

Fig 1. Histogram of mortality rates of 51 states as of April 13, 2020.

- Design a supervised learning workflow that leverages and compares 8 types of supervised learning algorithms.
- To help identify target patients and inform treatment and prevention.

Methods

Measurements

- Mortality rate of COVID-19 by state as of April 13, 2020, and as of Aug 27, 2020
  - https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data, operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE)
- ✓ A binary outcome was derived as mortality rate increase vs. decrease as of Aug 27 since April 13.
- A total of 242 social-economic, behavioral healthcare risks: from nationally representative surveys.
- ✓ the data have been aggregated at the state level, appearing as % of the state population, and reflecting the risks prior to the outbreak.
  - The Behavioral Risk Factors Surveillance System (BRFSS) Prevalence Data 2018[4]: 19 domains covering demographics, health-related risk behaviors, chronic health conditions, and use of preventive services
  - The US County Health Rankings & Roadmaps 2020[5]: health factors, such as length and quality of life, health behaviors, clinical care, social economic factors, and physical environment
  - The US Hospital Capacity of 2020[6]: state-level hospital bed occupancy rate and ICU bed occupancy rate

Supervised Learning Strategy

250 random splits of training and testing sets with 70:30 (36 vs. 15 states);
at each split, training/testing pair run by 8 types of learning algorithms

Train: 3-fold CV to determine optimal hyperparameters	Test: external validation

Each of 8 types of algorithms yielded an optimal model that gave the smallest test error of all splits

Full validation: apply the optimal model of each type to all 51-state data and yield full test error;

Final model leveraged external and full validation with small test and full errors;

Relative importance to determine key risks.

Results

- Distribution of mortality rate by state as of April 13, 2020: mean ± std was 3.01 ± 1.24, and [min, max] = [0, 6.1] (Fig 1).
- 18 states had mortality rate increased vs. 33 states decreased from April 13 to Aug 27, 2020.
- Of 242 predictors, 8 key risks were identified to impact mortality rate, 10 to mortality increase, by relative importance.
  - To predict mortality rate, Random Forest (RF) model appeared to yield small and tight MSEs from the full dataset validation, and meantime the MSEs from the external validation using the test set was also small (Table 1). Top features identified from the RF included % of population experienced sever housing cost burden, % of veterans (Fig 2), % of blacks, % of population aged 45 – 54 years, and % of adults (18-64yr)'s healthcare coverage.
  - In terms of explaining the mortality rate increase vs. decrease from April to Aug, 2020, the Gradient Boosting Machine (GBM) seems to provide with a better prediction (Table2& Fig3).

Top 2 features from the optimal GBM included % population who had a routine checkup 5+ years ago, and % population aged 45 – 54 years old.

Table 1. Models in Predicting Mortality Rate

Types of Learning	Hyperparameters from 3-fold CV (Training Set)	Mean MSE (±std) (Test Set)	Mean MSE (±std) (Full Set)
Lasso	alpha=0.12; tol=0.001	1.65±0.56	1.49±0.09
Ridge	alpha=1.05; tol=0.001	2.28±0.76	0.88±0.21
K-nearest Neighbors	algorithm=ball_tree; n_neighbors=5	1.76±0.50	1.32±0.09
SVM*	C=0.6 ;loss=squared_ epsilon_insensitive	3.40±2.44	3.04±2.27
CART**	crit=mse; max_depth= 2; max_feature= log2	2.16±0.74	1.30±0.23
Gradient Boosting Machine (GBM)	learning_rate=0.1 ; max_depth= 4; min_samples_leaf=3 ; max_features= log2	1.73±0.47	0.66±0.20
Random Forest (RF)	max_depth=4 ; min_samples_leaf=2; max_features=sqrt; min_samples_split=3	1.62±0.46	0.84±0.13
Multilayer Perceptron	activation=tanh; hidden_layer=(50,)*10	1.73±0.60	1.59 ±0.31

*Support Vector Machine. **Classification and Regression Tree.

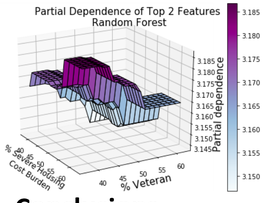

Fig 2. Partial dependence of top 2 features from RF. The mortality rates of the states with >50% of population had severe housing cost burden and veterans < 50% of the population were much higher than other states.

Conclusions

- Social-economic, behavioral and healthcare factors were collected prior to the pandemic, and can serve as potential predictors to mortality rate and mortality change by COVID-19. The key risks identified were consistent with others findings with traditional methods[7] .
- Comparing a diverse set of models (8 types) allows to avoid making assumptions about data.
- 2 categories of models (tree-based models, GBM or RF, and regularized regressions, Ridge or Lasso work well on a dataset with a small number of observations and a large number of predictors.

Table 2. Models in Predicting Mortality Rate Increase

Model Types	Mean Log Loss (±std) (Test Set)	Mean Log Loss (±std) (Full Set)	ROC AUC (mean±std) (Full Set)
Lasso	0.69±0.21	0.40±0.04	0.90±0.04
Ridge	0.98±0.36	0.35±0.11	0.91±0.04
KNN*	2.32±2.22	1.06±0.67	0.74±0.03
SVM	0.71 ±0.15	0.68±0.11	0.48±0.19
CART	3.05±2.94	1.26±0.89	0.73±0.07
GBM	0.63±0.20	0.31±0.06	0.96±0.03
RF	0.57±0.08	0.34±0.03	0.95±0.03
MLP**	0.67±0.07	0.64±0.03	0.59±0.11

*K-nearest neighbors **Multilayer Perceptron (a type of Neural Network - NN)

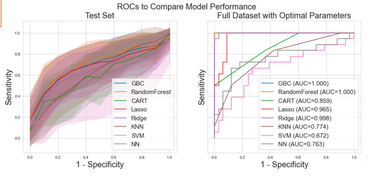

Fig 3. ROCs of 8 types of algorithms for test (left) and full validation (right) procedures.

Reference

1. U.S. Food & Drug Administration, Artificial intelligence and machine learning in software as a medical device, https://www.fda.gov/medical-devices/.
2. Hastie T, Tibshirani R, Friedman J, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed, 12th printing. New York, NY, Springer, 2017.
3. Centers for Disease Control and Prevention: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html
4. CDC BRFSS: https://www.cdc.gov/brfss/
5. https://www.countyhealthrankings.org/
6. Harvard Global Health Institute: https://globalepidemics.org/hospital-capacity/
7. Yehia B.R., et.al.:https://jamanetwork.com/journals/jamanetworkopen/article-abstract/2769387

Acknowledgements

*Matthew Ye is a sophomore at UC Berkeley, and did a summer internship at Translational Medicine & Global Biometrics and Data Sciences Divisions, Bristol Myers Squibb. The work was supported by the organizations of his internship, and thanks for the support.