RISK IDENTIFICATION AND PREDICTION FOR LONG COVID

Emma Jin, Elliot Kim, Victoria Song
CPSC 66 (Fall 2021)

INTRODUCTION

Long COVID: persistence of symptoms post-COVID-19

- Our dataset: Headaches, Fatigue, Shortness of breath, Anxiety, Heart problems, Muscle aches, Dizziness

Goals: build a classifier for long COVID and identify risk factors of long COVID

- Gender: Women have 2x likelihood compared to men (Nabavi, 2020)
- Age: Mean age of long COVID patients = mean of non-long COVID patients + 4 (Nabavi, 2020)
- Comorbidities (Osmanov et al., 2021)

Dataset: COVID-19 Fall 2020 & Winter 2021 Community Supplement, MCBS (Medicare Current Beneficiary Survey)

5 Most Common Symptoms
(López-León et al., 2021)

Fatigue (58%)
Headache (44%)
Attention disorder (27%)
Hair loss (25%)
Shortness of breath (24%)

Prevalence (Carfì et al., 2020)

35% of COVID outpatients
87% of inpatients

- *1-2 symptoms: 32%*
- *3+ symptoms: 55%*

TABLE OF CONTENTS

01

DATA PRE-PROCESSING

Feature selection and extraction

02

BUILDING OUR MODEL

Build a classifier that accurately classifies whether a patient has long COVID

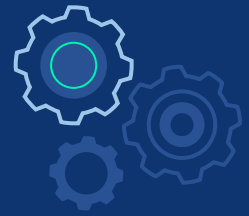
03

ANALYZING RISK FACTORS

Identify the most defining risk factors for long COVID

01

PRE-PROCESSING DATA



DIMENSIONALITY REDUCTION

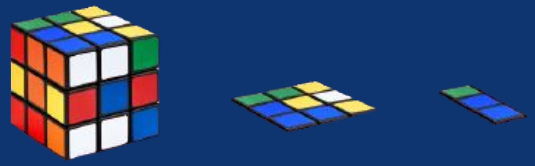
FEATURE SELECTION

Discard noisy or useless features

For example...

- ➔ The user_id of the subject
- ➔ The week in which the subject completed the survey

We also discarded features that are shown to have trivial importance



FEATURE EXTRACTION

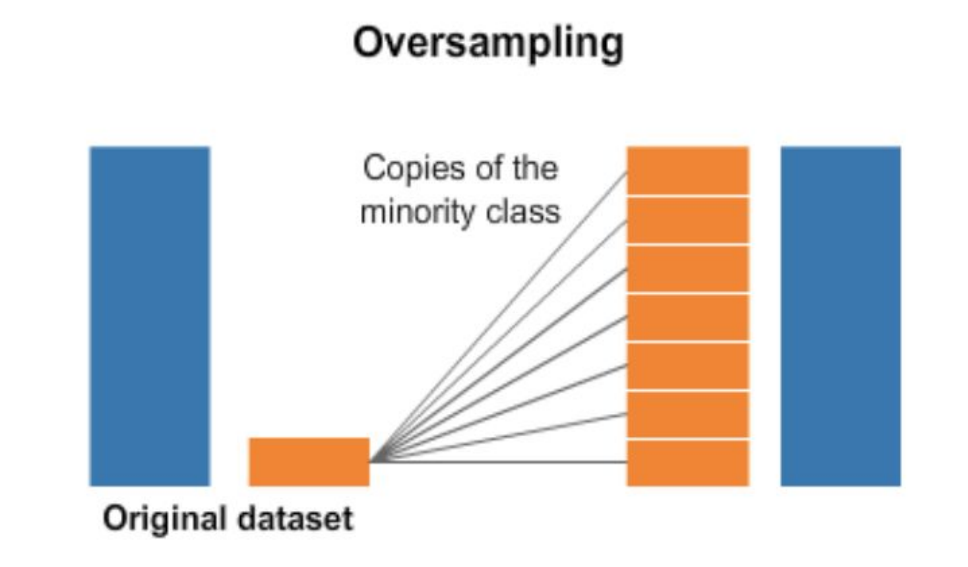
Produce a new set of features from the old ones

For example...

- Do experts recommend hand-washing?
- ... wearing masks?
- ... avoiding gatherings?
- ... staying at home?
- ... medical attention for short breath?

→ Knowledge of COVID measures recommended by experts

RESAMPLING THE DATA



UPSAMPLING/ OVERSAMPLING

It's important to note that we're **only** performing upsampling on the **training set**!

Our testing set stays the same

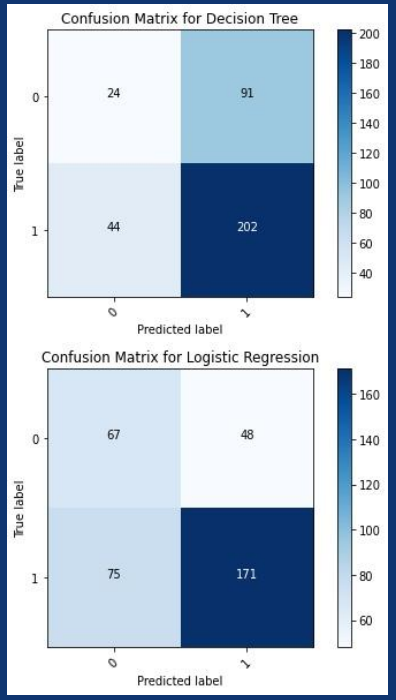
02

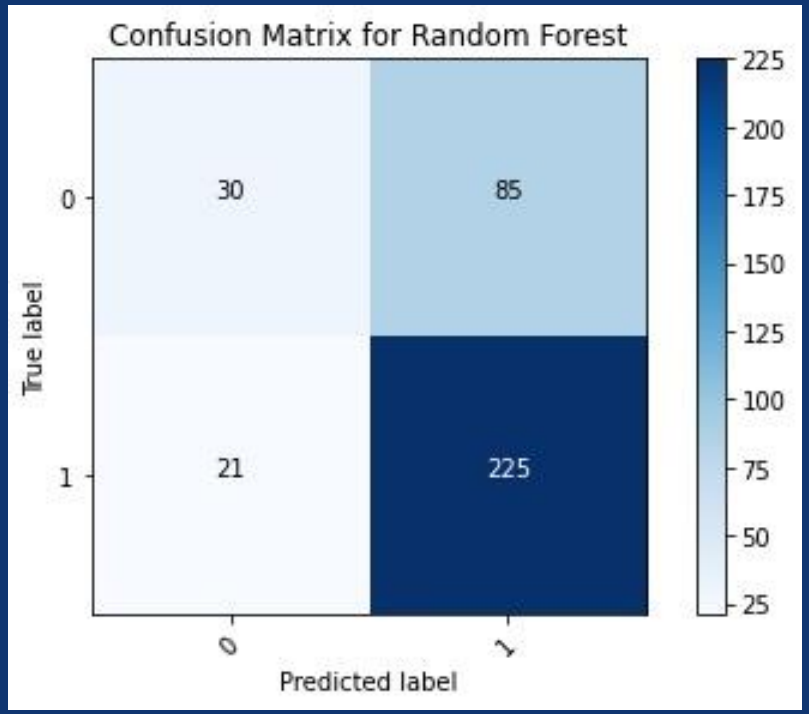
BUILDING OUR MODEL

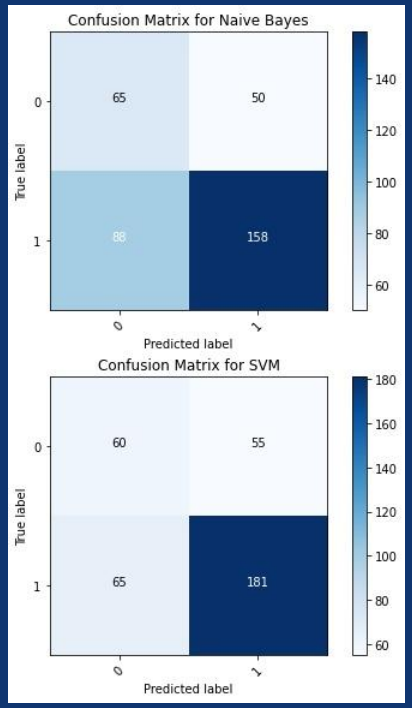
TABLE: ACCURACY OF DIFFERENT MODELS

	INITIAL MODEL ACCURACY	AVG TUNED TESTING ACCURACY
Logistic Regression	0.660	0.686
Decision Tree	0.626	0.624
Random Forest	**0.723**	**0.720**
SVM	0.668	0.693
Naive Bayes	0.618	0.654

CONFUSION MATRICES AFTER TUNING



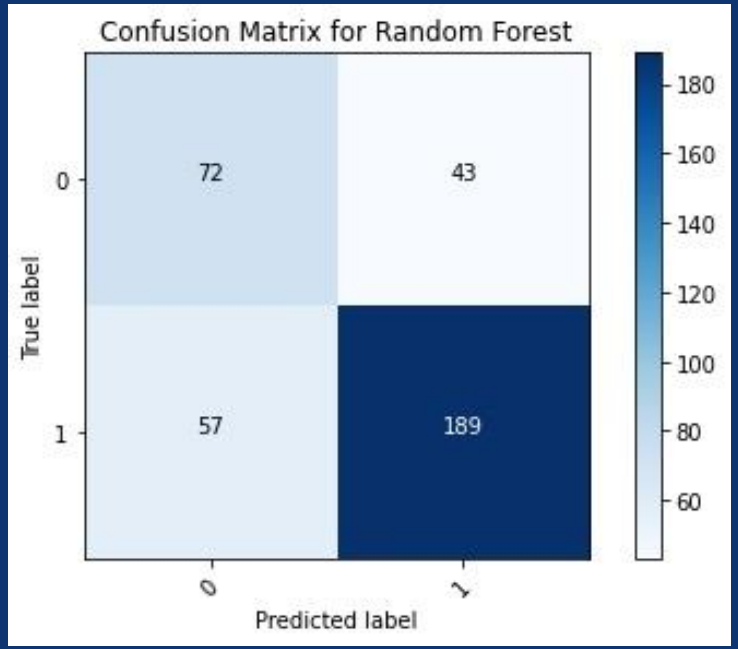




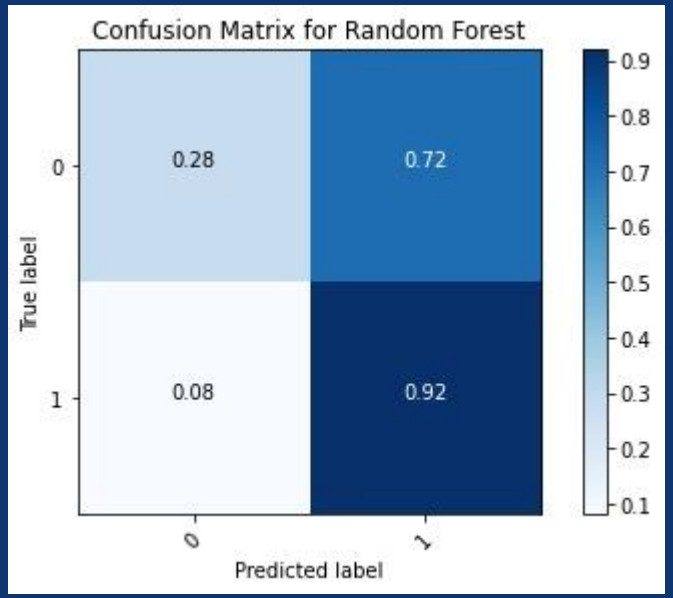
TUNING ON PRECISION

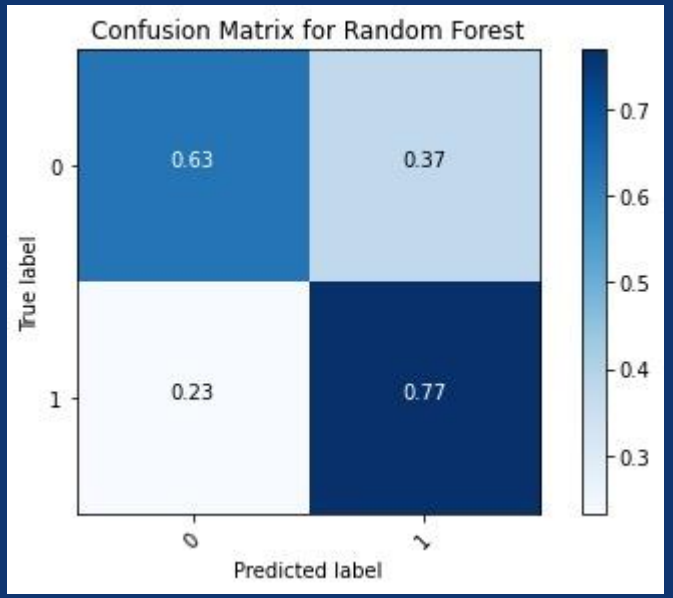
- number of estimations = 75, max depth = 5, max leaf nodes = 15, max leaf nodes = 15, max features = 50% and min samples leaf = 25

- Performing a hold-out validation on this model yields accuracy = 0.723, precision = 0.815, recall = 0.768, and f1-score= 0.791



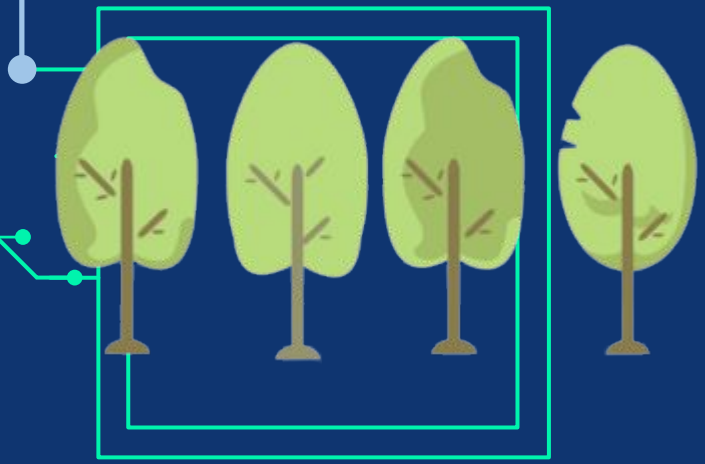
NORMALIZED CONFUSION MATRIX





03

ANALYZING RISK FACTORS



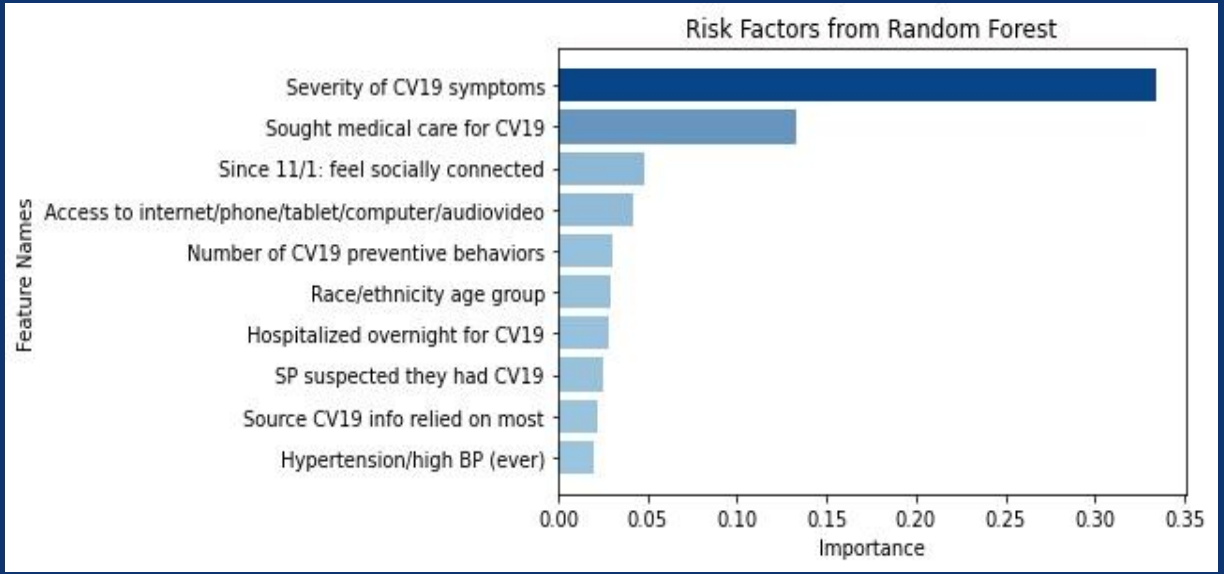
GINI IMPORTANCE FOR RANDOM FORESTS

The Gini importance for a feature θ is defined as:

The total decrease in node impurity when θ is chosen to split the node , averaged over all trees in the forest

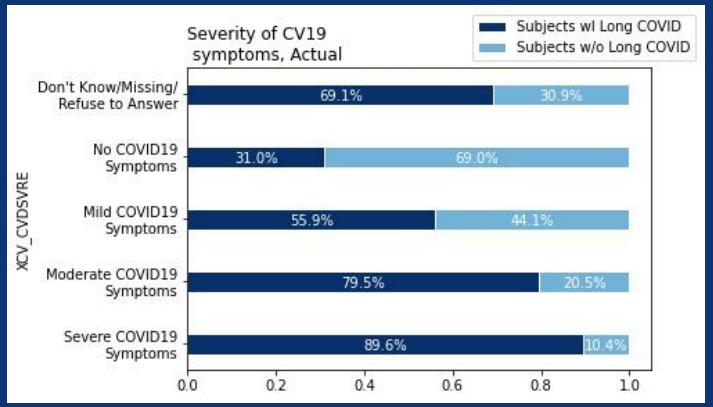
IDENTIFYING RISK FACTORS

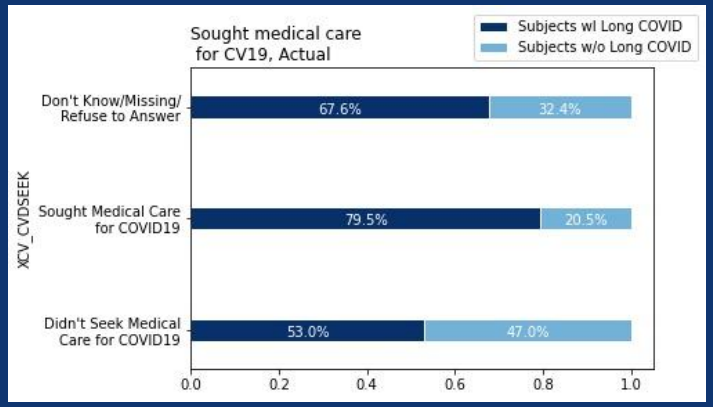
1. SEVERITY OF CV19 SYMPTOMS
2. SOUGHT MEDICAL CARE FOR CV19

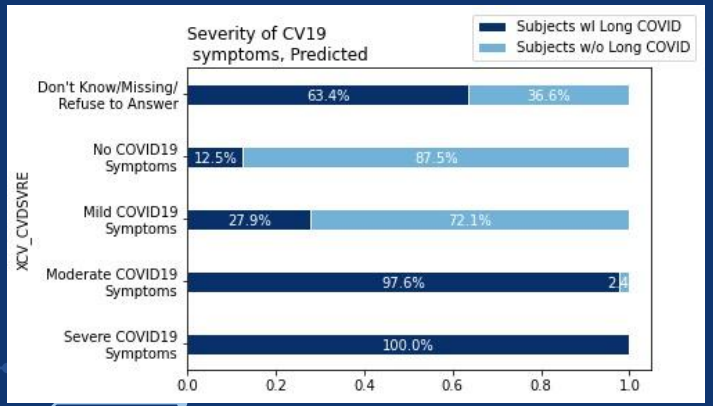


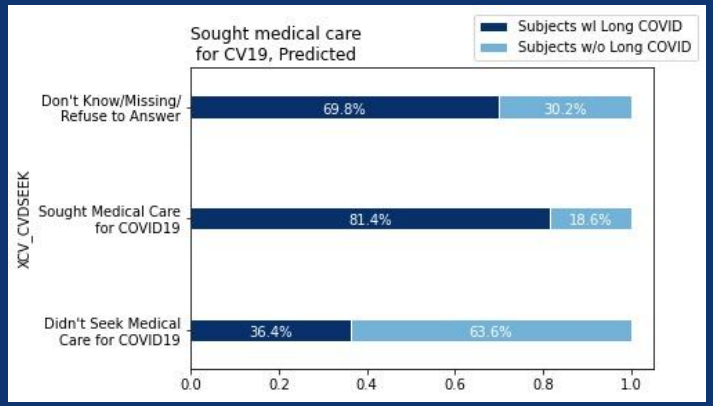
PERCENTAGES BY VALUE

PERCENTAGES BY VALUE









CONCLUSION

- Random Forest Classifier with 0.723 accuracy and 0.815 precision
- Most prominent risk factors: "Severity of COVID-19 Symptoms" and "Sought medical care for COVID-19"
- More severe COVID symptoms ⇒ greater likelihood of long COVID
- Patients who sought medical care for COVID are more likely to have long COVID
- Patients can still get long COVID with mild or even no COVID symptoms.

CONCLUSION

Implications

- Reinforce the importance of existing research dedicated to preventing and minimizing severe COVID-19.
- Indicate potential need for public health messaging to emphasize risk of long COVID even for patients with no or mild COVID symptoms

Future Directions

- Develop more ways to alleviate severe COVID
- Reconsider definition of “fully recovered from COVID-19”

REFERENCES

- Nisar, Riddhi. "(Visually) Interpreting the Confusion-Matrix: | by Riddhi Nisar | Analytics Vidhya | Medium." *Medium*, Analytics Vidhya, 1 Nov. 2020, https://medium.com/analytics-vidhya/visually-interpreting-the-confusion-matrix-787a70b65678.
- Chen, Denise. "Statistical Learning: Data Sampling & Resampling." Medium, Towards Data Science, 12 Apr. 2020, https://towardsdatascience.com/statistical-learning-ii-data-sampling-resampling-93a0208d6bb8.
- "320000+ Environmental Cartoons Images, HD Pictures and Stock Photos for Free Download." LovePik, https://lovepik.com/images/environmental-cartoons.html.
- Mahendrakumaran, Kayathiri. "Dealing with Imbalanced Data." Medium, தழலி, 20 Dec. 2020, https://medium.com/%E0%AE%A4%E0%AE%B4%E0%AE%B2%E0%AE%BF/dealing-with-imbalanced-data-aca93c421fff.
- López-León, Sandra, et al. "More than 50 Long-Term Effects of COVID-19: A Systematic Review and Meta-Analysis." SSRN Electronic Journal, 20 Jan. 2021, https://doi.org/10.2139/ssrn.3769978.
- Osmanov, Ismail M, et al. "Risk Factors for Long Covid in Previously Hospitalised Children Using the ISARIC Global Follow-up Protocol: A Prospective Cohort Study." European Respiratory Journal, 2021, p. 2101341., https://doi.org/10.1183/13993003.01341-2021.
- Nabavi, Nikki. "Long Covid: How to Define It and How to Manage It." BMJ, 2020, p. m3489., https://doi.org/10.1136/bmj.m3489.
- Carfì, Angelo, et al. "Persistent Symptoms in Patients after Acute COVID-19." JAMA, vol. 324, no. 6, 11 July 2020, p. 603., https://doi.org/10.1001/jama.2020.12603.

THANK YOU!