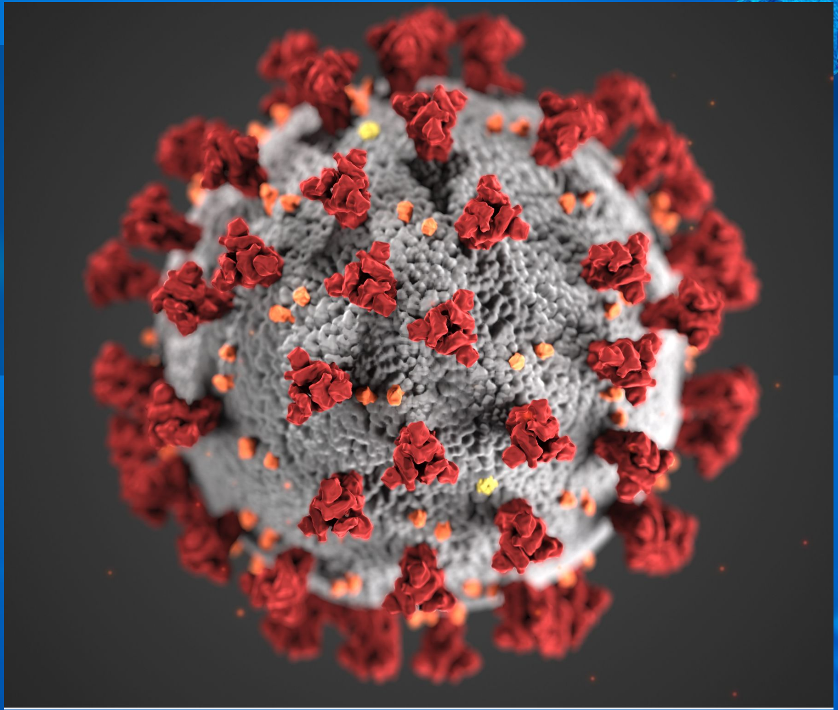
COVID 19: Symptoms and Predictions

By: David Cortes, Nao Kawakami, Peter Yonka, Ryan Scribner



Problem Statement:

Utilizing the Coronavirus Disease 2019 (COVID-19) Clinical Data Repository from Carbon Health, we will examine the influence of individual symptoms on whether a COVID-19 test will be positive as well as build a purely predictive model.

We hope that the understanding gained will help frontline medical works at Carbon Health with limited time and resources to triage cases and determine initial steps in creating treatment plans.

Problem Statement: Part 1: Influence of Symptoms

Interpretability of feature influence is of the utmost importance for this section of our problem statement, so we broke down the symptoms into two different groups:

1. Patient reported features (able to be collected prior to interaction with patient via an app or questionnaire)
2. Patient report features combined with clinically collected/assessed features (require availability of resources or interaction with patient).

Each of these datasets were run through a logistic regression optimizing for true positives and analyzed at the 99% confidence level (alpha = 0.01).

Problem Statement: Part 2: Predictive Model

The dataset will be cleaned and modeled using a suite of classification models and a voting classifier to maximize predictive power.

Use cases for the predictive model would be for prioritizing patient care if rapid testing is not available or to prioritize testing of samples with a higher likelihood of testing positive.

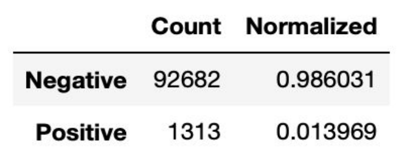
Models:

We designed two different data sets:

1. Identify which symptoms are the most important to accurately assessing patients for COVID 19.
2. Can we accurately predict if someone will test positive based on their symptoms.

Initial Challenges:

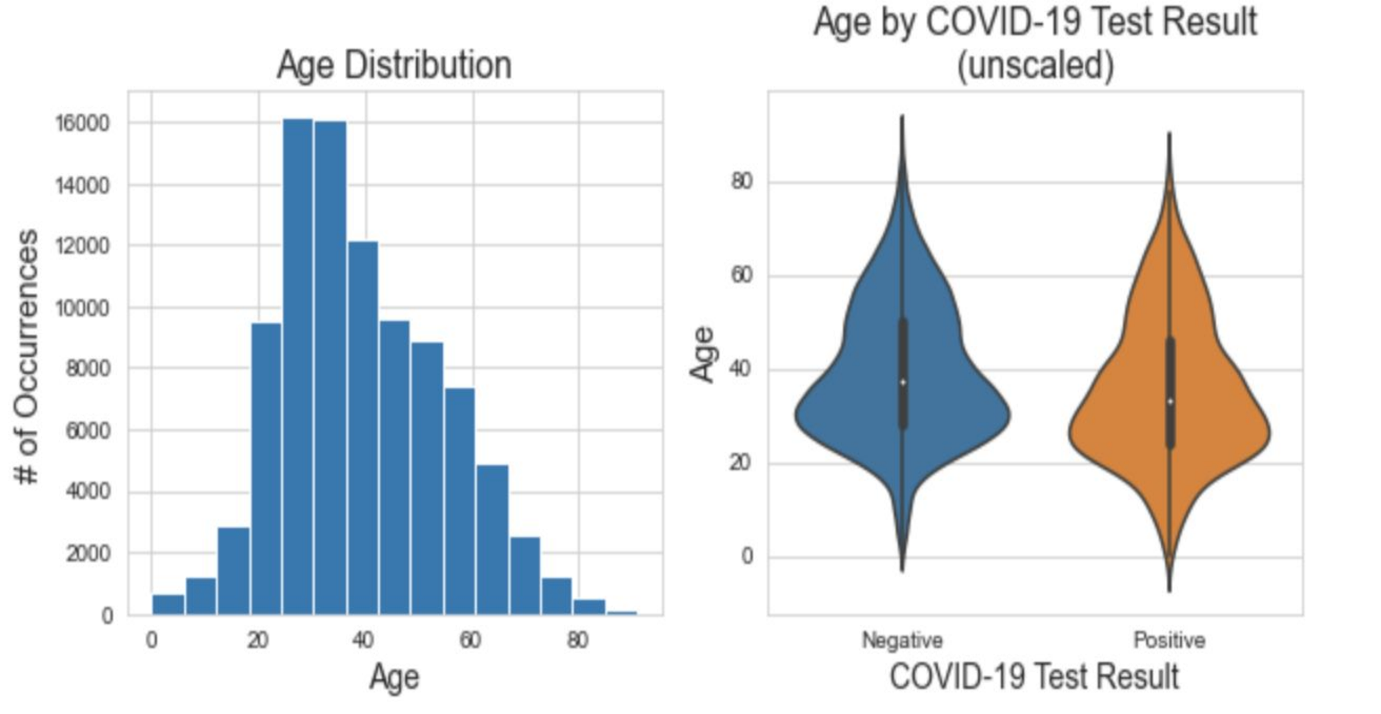
- There is an overwhelming amount of negative results compared to positive
- How do we clean the data without losing to many positive results?
- LabCorp and Quest Diagnostics reports a false negative rate between 3-15%. The longer the duration, the fewer the false negatives.

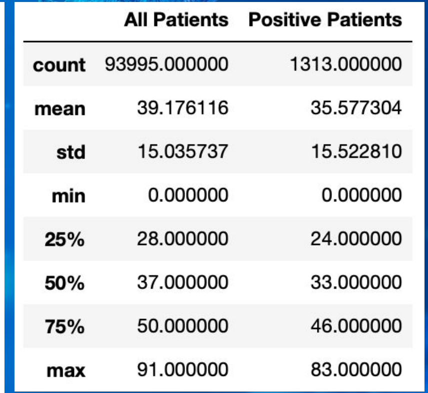


	Count	Normalized
Negative	92682	0.986031
Positive	1313	0.013969

Important Facts:

It is important to keep in mind who we are testing! Since this clinic is using an application/technology to initially test patients what does this mean?





	All Patients	Positive Patients
count	93995.000000	1313.000000
mean	39.176116	35.577304
std	15.035737	15.522810
min	0.000000	0.000000
25%	28.000000	24.000000
50%	37.000000	33.000000
75%	50.000000	46.000000
max	91.000000	83.000000

Important Symptoms/Vitals:

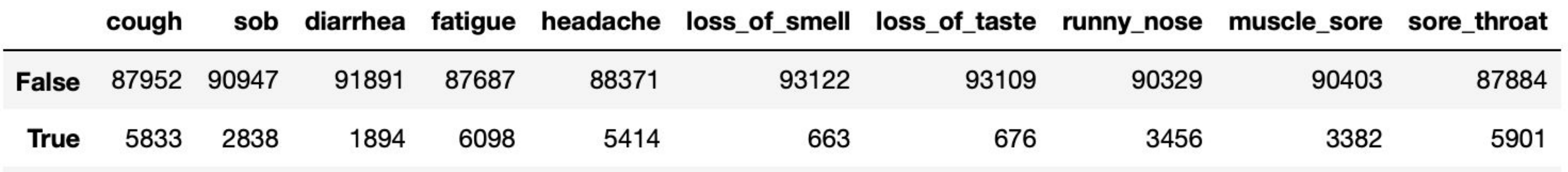
Patient reported symptoms:

- Days since symptom onset
- Cough/cough severity
- Fever
- Shortness of breath/sob severity
- Diarrhea
- Fatigue
- Headache
- Loss of smell
- Loss of taste
- Runny nose
- Sore muscles
- Sore throat

Clinically collected vitals:

- Temperature
- Pulse
- Systolic blood pressure
- Diastolic blood pressure
- Respiratory rate
- Oxygen saturation

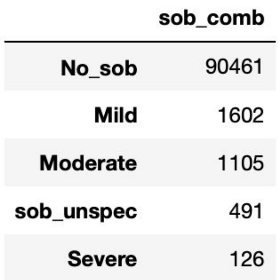
Model 1 EDA (Symptoms):



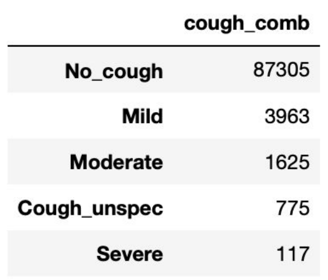
	cough	sob	diarrhea	fatigue	headache	loss_of_smell	loss_of_taste	runny_nose	muscle_sore	sore_throat
False	87952	90947	91891	87687	88371	93122	93109	90329	90403	87884
True	5833	2838	1894	6098	5414	663	676	3456	3382	5901

EDA:

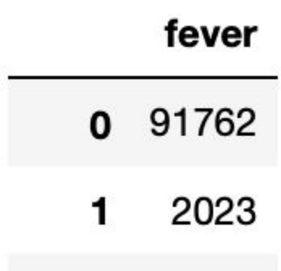
- Isolate columns of interest
- Combine similar columns
- Drop null values
- Assign numeric values



	sob_comb
No_sob	90461
Mild	1602
Moderate	1105
sob_unspec	491
Severe	126

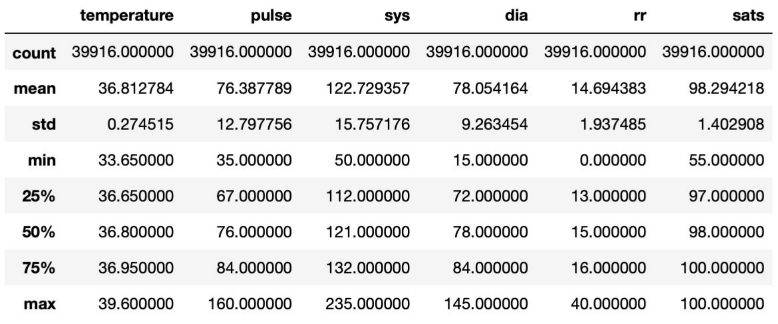


	cough_comb
No_cough	87305
Mild	3963
Moderate	1625
Cough_unspec	775
Severe	117

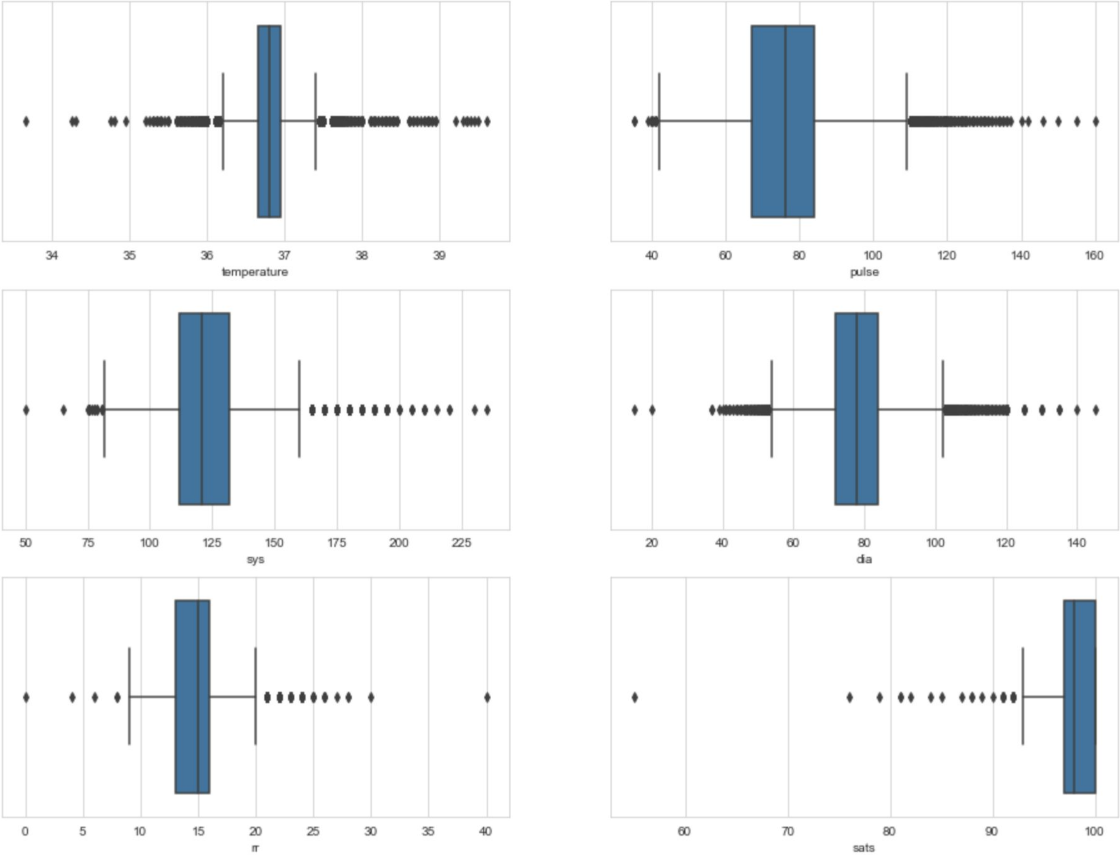


	fever
0	91762
1	2023

Model 1 EDA Vitals:



	temperature	pulse	sys	dia	rr	sats
count	39916.000000	39916.000000	39916.000000	39916.000000	39916.000000	39916.000000
mean	36.812784	76.387789	122.729357	78.054164	14.694383	98.294218
std	0.274515	12.797756	15.757176	9.263454	1.937485	1.402908
min	33.650000	35.000000	50.000000	15.000000	0.000000	55.000000
25%	36.650000	67.000000	112.000000	72.000000	13.000000	97.000000
50%	36.800000	76.000000	121.000000	78.000000	15.000000	98.000000
75%	36.950000	84.000000	132.000000	84.000000	16.000000	100.000000
max	39.600000	160.000000	235.000000	145.000000	40.000000	100.000000



EDA:

- Drop null values
- Normalize

Model 1: Baseline Scoring

Logistic Regression

Mean cross-val accuracy score	0.986
Mean cross-val recall score	0.023

Logistic Regression (class_weight=’balanced’)

Mean cross-val accuracy score	0.872
Mean cross-val recall score	**0.521**

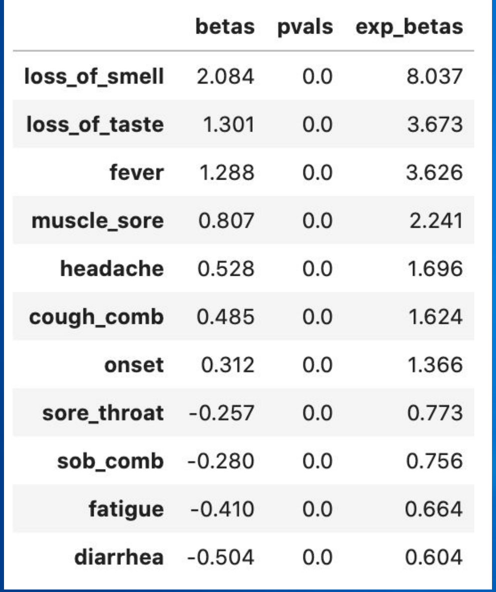
Model 1: Balancing Act

Enter imbalanced-learn:

- RandomOverSampler
- RandomUnderSampler
- imbalanced-learn Pipeline

Grid Search Logistic Regression Model Scores	
Training accuracy score	0.868
Testing accuracy score	0.868
Training recall score	0.526
Testing recall score	0.537

Model 1: Patient Reported Symptoms Results

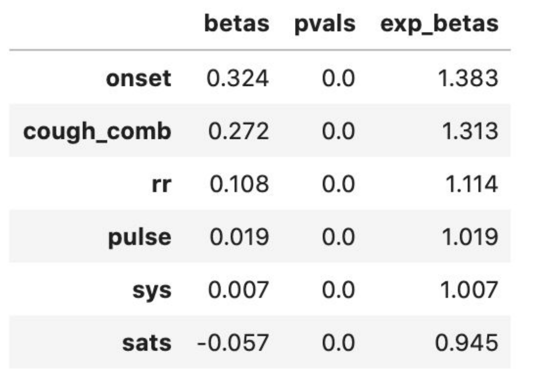


	betas	pvals	exp_betas
loss_of_smell	2.084	0.0	8.037
loss_of_taste	1.301	0.0	3.673
fever	1.288	0.0	3.626
muscle_sore	0.807	0.0	2.241
headache	0.528	0.0	1.696
cough_comb	0.485	0.0	1.624
onset	0.312	0.0	1.366
sore_throat	-0.257	0.0	0.773
sob_comb	-0.280	0.0	0.756
fatigue	-0.410	0.0	0.664
diarrhea	-0.504	0.0	0.604

Let's interpret the first three features:

- If a patient reports the symptom loss of smell, the patient is 8.037 times as likely to test positive for COVID-19, all else being held constant.
- If a patient reports the symptom loss of taste, the patient is 3.673 times as likely to test positive for COVID-19, all else being held constant.
- If a patient reports the symptom fever, the patient is 3.626 times as likely to test positive for COVID-19, all else being held constant.

Model 1: Patient / Clinical Combined Results



	betas	pvals	exp_betas
onset	0.324	0.0	1.383
cough_comb	0.272	0.0	1.313
rr	0.108	0.0	1.114
pulse	0.019	0.0	1.019
sys	0.007	0.0	1.007
sats	-0.057	0.0	0.945

Let's interpret the first three features:

- For every 1-unit increase in the onset measurement, the patient is 1.383 times as likely to test positive for COVID-19, all else being held constant.
- For every 1-unit increase in cough rating, the patient is 1.313 times as likely to test positive for COVID-19, all else being held constant.
- For every additional breath per minute in respiratory rate, the patient is 1.114 times as likely to test positive for COVID-19, all else being held constant.

Model 1: Summary and Recommendations

We were able to offset the imbalanced classes by utilizing over/undersampling techniques only to a point, especially when optimizing for the true positive rate. However, the models that were fit do provide some insight into the influence of symptom features in relation to a patient testing positive on a COVID-19 test.

- Utilize the odds for each feature to weigh questionnaire responses when computing testing priority recommendations for the patient and clinic resource management
- Update and assess weights weekly based on new testing data to revise as needed
- Share the results with frontline staff
  - Does this match what they are seeing?
  - Are there other things you should be looking for?
- Check on the data integrity practices surrounding feature measurements of your clinical data
  - A lot of potential outlier data

Model 1: Next Steps

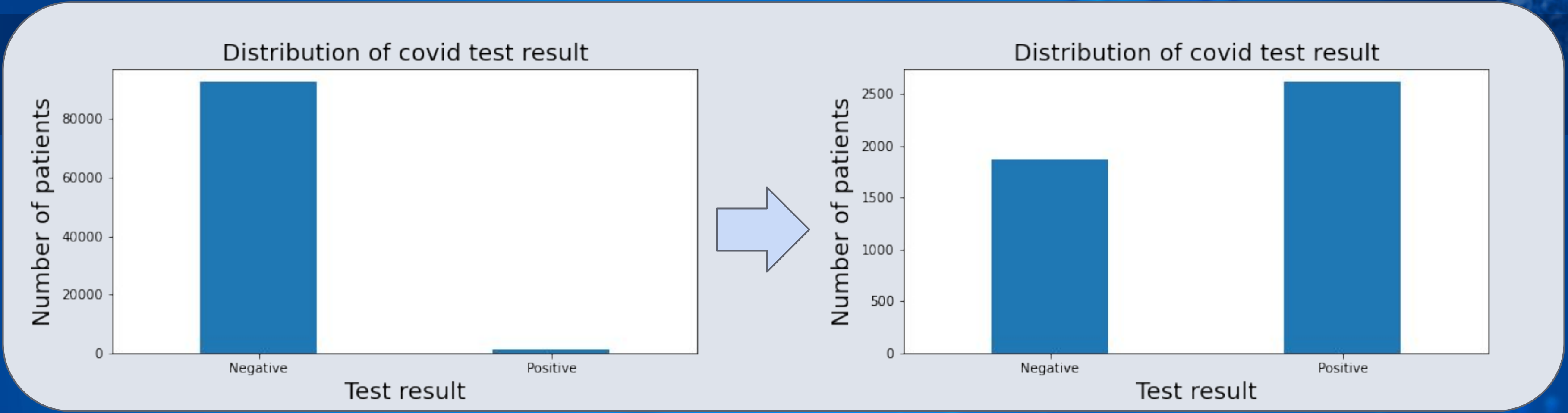
- Gather more data
  - Any way to pool with other clinics?
  - Access to more positive test results
- Test alternative over/undersampling methodologies
- Clarity from Carbon Health on clinically collected data for further cleaning and correction
- Dashboard with symptom metrics updated with new data as it is available

Model 2: Goals

For frontline medical workers, this model helps to:

- Pre-screen and prioritize potential patients
- Group them based on the result and avoid infection onsite

Model 2: Predictive Model EDA



1. Separate to negative and positive
2. Drop all missing values from negative
3. Fill missing values of positive with random values
    a. Binary columns → Binomial distribution
    b. Other columns → Normal distribution
4. Resample positive

Model 2 Results: Accuracy

Estimator	Train (%)	Test (%)
LogisticRegression	82	83
SVM	92	88
Keras Sequential	90	87
VotingClassifier	97	93

Model 2 Results: Accuracy

VotingClassifier parameter
AdaBoostClassifier (base_estimator=DecisionTreeClassifier(max_depth=2), n_estimators=125)
GradientBoostingClassifier(n_estimators=50)
DecisionTreeClassifier(max_depth=6)
KNeighborsClassifier(n_neighbors=3)
XGBClassifier

Train	Test
97	93

Model 2 Summary:

Although our predictive models are very accurate, they probably aren’t accurate enough for the medical field or a possible life and death determination.

Model 2: Future Steps:

- Tune hyperparameters
- Add more classifiers
- More samples of positive

References:

- https://carbonhealth.com/coronavirus
- Coronavirus Disease 2019 (COVID-19) Clinical Data Repository: https://github.com/mdcollab/covidclinicaldata