ALBERT EINSTEIN
SOCIEDADE BENEFICENTE ISRAELITA BRASILEIRA

TO: Dr. Cláudio Lottenberg, President

FROM: Patrick Curran, Senior Data Analyst

DATE: May 17, 2020

SUBJECT: Diagnosis of COVID-19 and Its Clinical Spectrum

As cases of COVID-19 continue to increase across Brazil, and the availability of tests continues to be a problem, extreme efforts must be undertaken to quickly identify positive cases. Our ability to identify positive cases will allow for rapid self-isolation measures in the hopes of slowing the spread of the virus. Since my update two weeks ago, two additional models were created in the hopes of better identifying positive COVID-19 patients. A polynomial support vector machine and a logistic regression model were developed. However, the linear support vector machine model still does the best job of identifying positive patients.

As of May 15, 2020, Brazil has reported over 220,000 confirmed cases, which is an increase of 150,000 cases in the last two weeks since my last memo. These new cases have also resulted in over 15,000 deaths, an increase of 10,000 deaths in the last two weeks. Since Brazil is only testing those who end up in the hospital, these numbers are most likely an underestimate due to the lack of testing. As President Jair Bolsonaro continues to downplay the severity of the virus, mitigation strategies are being left up to the governors of the individual states in Brazil. Therefore, the actions we take as a hospital are even more important.

The overwhelming objective of the models I developed is to help clinicians identify as many positive cases of COVID-19 with the limited resources of the hospital. When analyzing the top fourteen models (see Attachment 1), the greatest consideration was still given to the sensitivity (recall) of the model. In other words, how well the model performs at identifying positive cases. Under the assumption that testing will remain limited, or non-existent, the model has to be a reliable companion in correctly identifying positive cases. This focus on maximizing recall will result in a decline in the specificity, or the true negative rate, meaning the model will produce more false positives. However, this is a trade-off we must continue to be willing to make with the limited resources currently available at the hospital. It is more tenable to tell a patient they are positive when they actually are not than it is to tell a patient they are not positive when they, in fact, are.

With research showing every person infected with COVID-19 will, on average, infect 2.6 other people, the chosen model was the linear support vector machine model with a recall rate of over 95%. The ROC Curve for this model is shown in Figure 1 below. The dotted blue line illustrates simply guessing a patient's status, so the closer the linear support vector machine model gets to the top left-hand corner of the graph, the better the model performs. This model produced the highest ROC AUC score at 0.870, which is the area under the orange curve. A

perfect model will have a score of 1.0, meaning the model can predict all true positives without producing a single false positive.

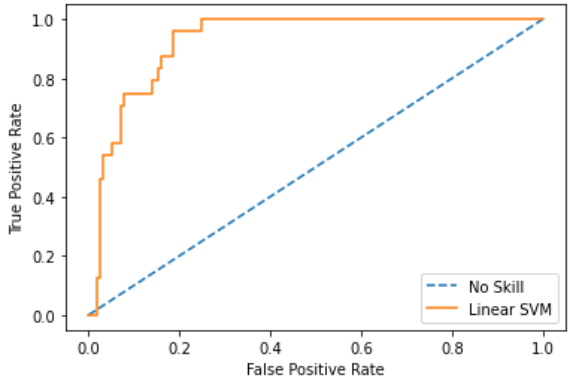

Figure 1: Linear Support Vector Machine ROC Curve

The other advantage of this model is the coefficients produced can help interpret which patient attributes effect a patient's status the most. In Figure 2, the most important predictors are shown. Those features displaying positive values indicate importance to a positive test, and those features displaying negative values indicate importance to a negative test. Not surprisingly, the most important predictor, by far, for a negative test is if a patient tests positive for another infection. Implementing process of elimination by using other tests that are in greater abundance is the first line of defense in determining a patient's status. If all of these tests come back negative, then blood work can be done. If no other infection can be determined, physicians should pay special attention to the seven features of a patient's blood work, as elevated values may be an indicator of the patient's status.

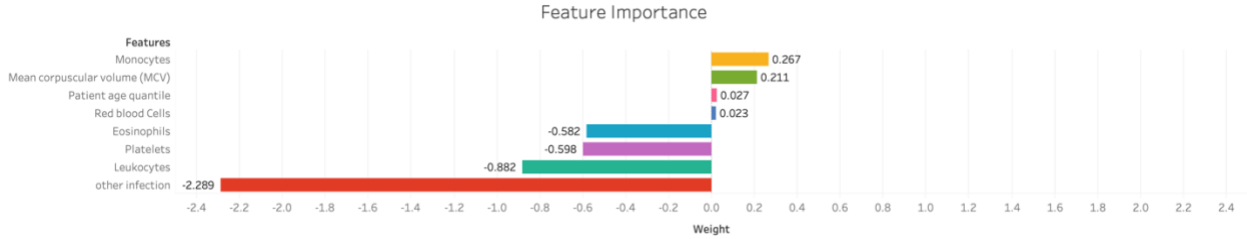

Figure 2: Linear Support Vector Machine Feature Importance

The chosen linear support vector machine model paired with patient information about whether he/she has been in contact with an individual(s) known to be positive for the virus can be an exceptional tool in the hands of our doctors in the fight against COVID-19.

Attachment 1: COVID-19 Diagnosis Model Metrics
Attachment 2: COVID-19 Diagnosis Model Confusion Matrices

COVID-19 Diagnosis Model Metrics
(Training Data: 418 patients, Testing Data: 180 patients)

Model	Accuracy	Recall	ROC AUC Score
kNN Model 2	0.878	0.292	0.630
DT Model 3	0.811	0.708	0.768
RF Model 2	0.911	0.583	0.772
ET Model 2	0.833	0.792	0.816
Linear SVC Model 3	0.806	0.958	0.870
RBF SVC Model 2	0.867	0.333	0.641
Poly SVC Model 2	0.872	0.500	0.715
ANN Model 2	0.878	0.625	0.771
SGD Model 2	0.894	0.542	0.745
GBC Model 2	0.922	0.542	0.761
Logit Model 3	0.811	0.917	0.856
BAG Model 3	0.900	0.750	0.837
AdaBoost Model 2	0.894	0.667	0.798
Stacking Model 1	0.811	0.917	0.856

COVID-19 Diagnosis Model Confusion Matrices
(Testing Data: 180 patients)

Model	Confusion Matrix		
		Predicted Negative	Predicted Positive
kNN Model 2	Actual Negative	151	5
	Actual Positive	17	7
DT Model 3	Actual Negative	129	27
	Actual Positive	7	17
RF Model 2	Actual Negative	150	6
	Actual Positive	10	14
ET Model 2	Actual Negative	131	25
	Actual Positive	5	19
Linear SVC Model 3	Actual Negative	122	34
	Actual Positive	1	23
RBF SVC Model 2	Actual Negative	148	8
	Actual Positive	16	8
Poly SVC Model 2	Actual Negative	145	11
	Actual Positive	12	12
ANN Model 2	Actual Negative	143	13
	Actual Positive	9	15
SGD Model 2	Actual Negative	148	8
	Actual Positive	11	13
GBC Model 2	Actual Negative	153	3
	Actual Positive	11	13
Logit Model 3	Actual Negative	124	32
	Actual Positive	2	22
BAG Model 3	Actual Negative	144	12
	Actual Positive	6	18
AdaBoost Model 2	Actual Negative	145	11
	Actual Positive	8	16
Stacking Model 1	Actual Negative	124	32
	Actual Positive	2	22