CovidCheck: A COVID-19 Predictor Based On Patient Symptoms

Agnim Agarwal

Abstract

In this report, we present a thorough analysis of FeverIQ's dataset consisting of symptom screener records captured into four preconfigured diagnostic vectors and results from a COVID-19 molecular diagnostic test. From our analysis, we present:

- Insight involving correlation between diagnostic vectors, date, and location both with test result and with other features, including a temporal and geographical analysis of specific features
- Prediction of COVID-19 PCR Test Result with up to 80% accuracy
- Interpretation of black box algorithm with plots of global feature importance and local feature importance for each prediction in testing dataset.

Using these insights, we quantify the impact of each symptom vector on a COVID-19 diagnostic test and demonstrate our machine learning pipeline's impact on future scientific, societal, and policy changes.

Table of Contents

Introduction

The ongoing COVID-19 pandemic has infected 39.6 million people and killed 1.11 million people worldwide. The United States has had one of the highest infection rates with 8.14 million people infected and 200,000 people killed as of October 2020.

FeverIQ provides one of the strongest routes of data collection with an online survey that securely circumvents the numerous regulatory and political barriers to collecting and sharing COVID-19 health data.

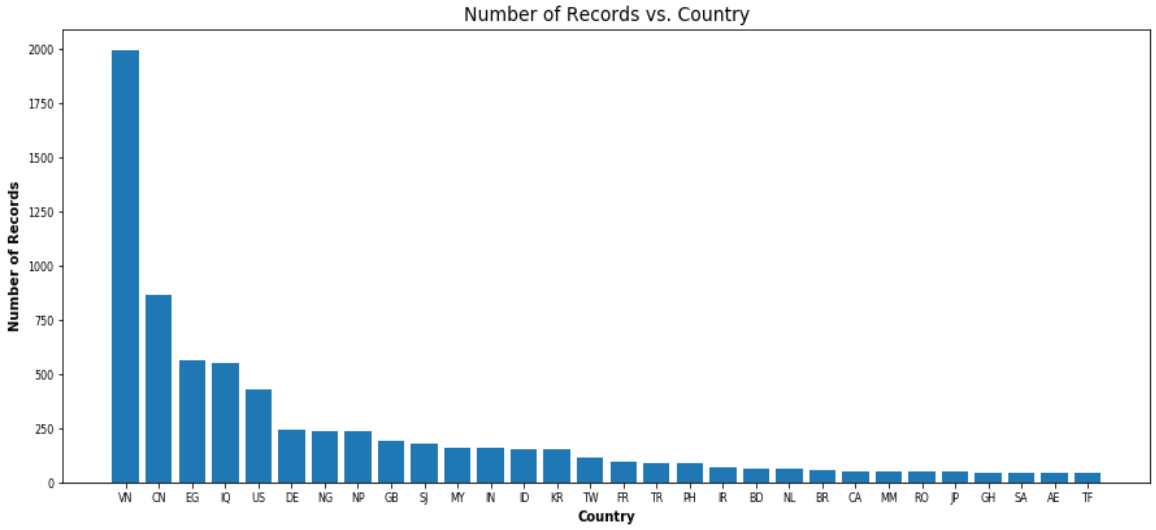
The dataset relies on four main diagnostic statistics “designed to capture the similarity of the user’s symptoms to four preconfigured diagnosis vectors”. From the dataset, it is in the best interests of epidemiological researchers, data scientists, and others to understand the implications of this dataset, including underlying patterns between certain symptoms and correlation between symptoms or sets of symptoms with a COVID-19 diagnostic test result.

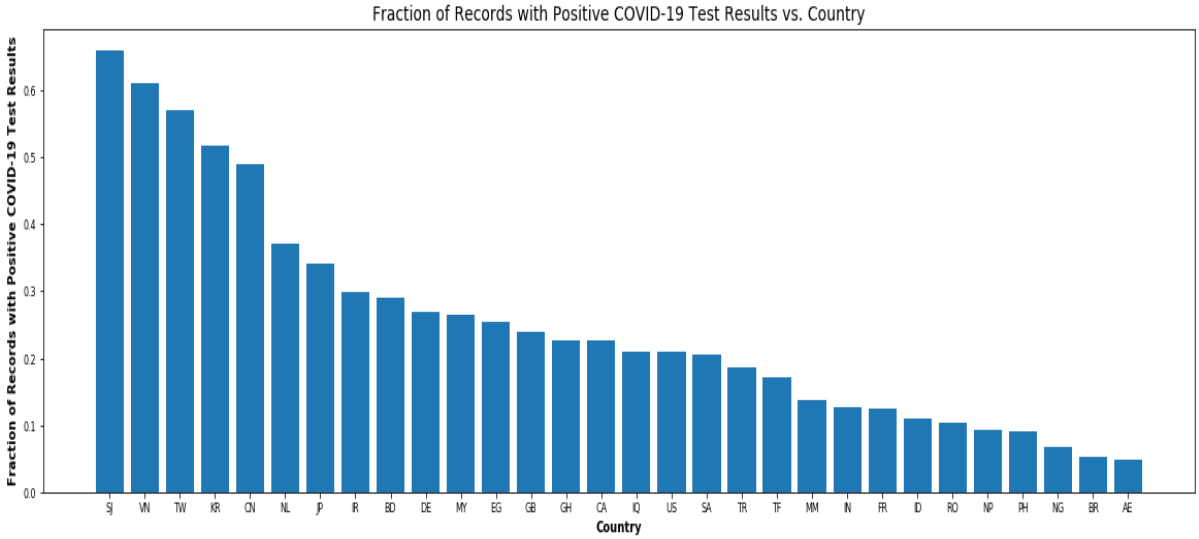
The dataset provided is a sample of records from May-June collected from FeverIQ, a privacy-preserving symptom screener for COVID-19, and results from a COVID-19 molecular diagnostic test.

Results

Feature Engineering and Correlation Analysis

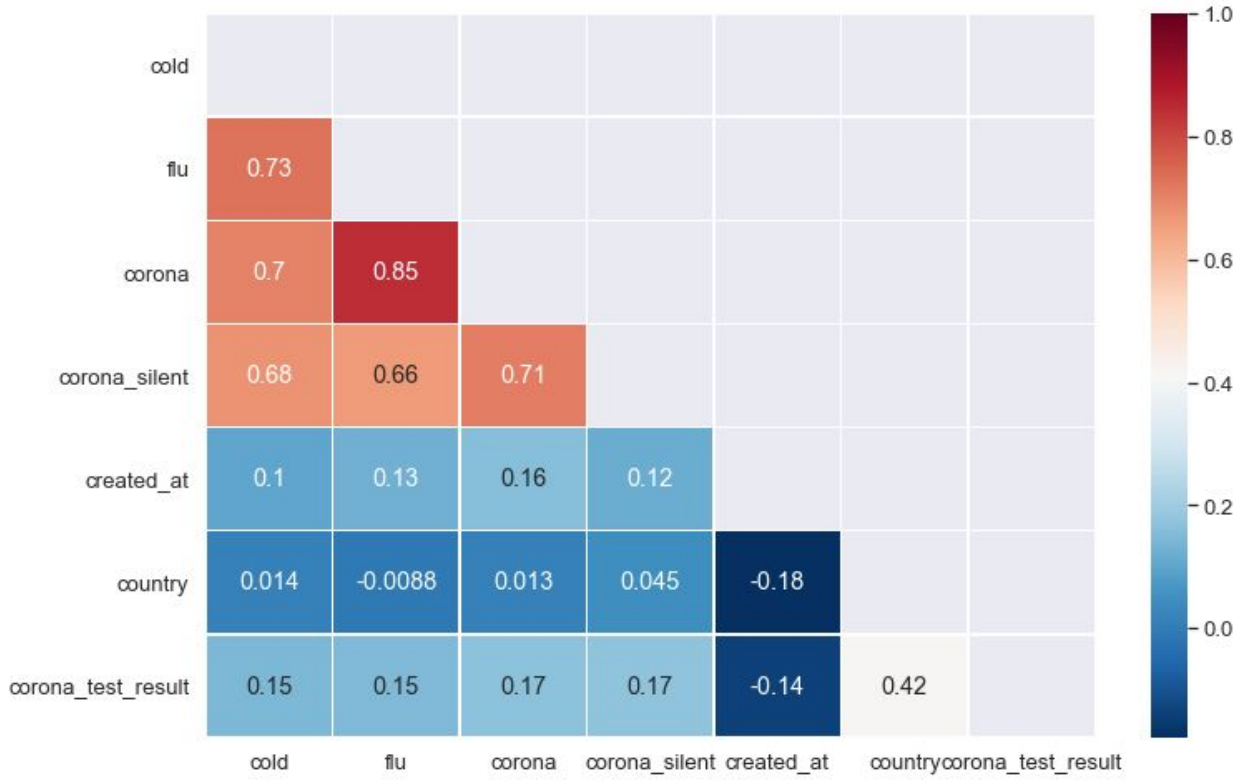
From the original dataset, we found it in our best interests to convert both the country code and date to a numerical format and create a tabular dataset. While there are many ways to place the country code feature in the context of COVID-19 prediction, we found it most useful to convert the country code to the fraction of positive cases in that country. For this problem, we calculated this from the number of records indicating a positive test result divided by the total number of records grouped by country. Because we also found there are some countries with very small sample sizes, we computed this value for the top 30 countries to generate a more accurate portrayal of each country's fraction of positive cases. Doing this creates a threshold of at least 40 positive cases country-wide moving forward. Below are the histograms for both number of records and fraction of positive cases vs. country.





For the date, we found it reasonable to convert to days since the first date recorded, May 6th, as this provides the date in a tabular format without losing any information.

From the engineered features, we computed a correlation matrix showing correlation coefficients between variables. Here, each cell in the table shows the correlation between two variables. The matrix was generated using scikit-learn's correlation matrix functionality and plotted with seaborn data visualization based on matplotlib.



Based on the correlation matrix, we find:

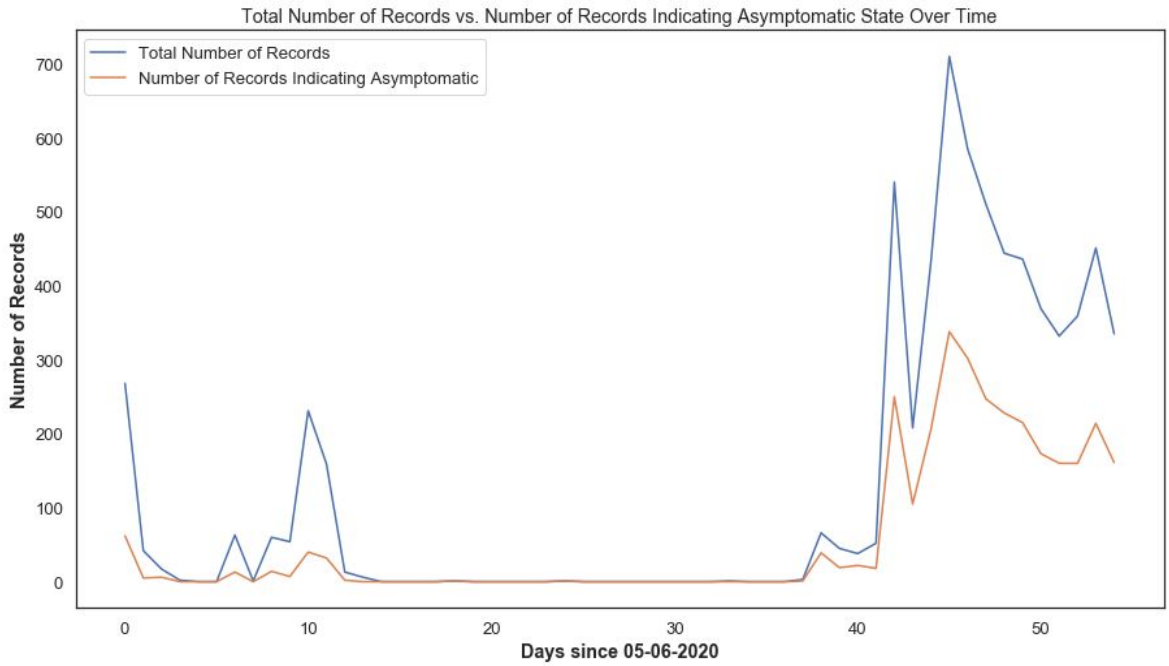
- The variables with the highest correlation are Corona and Flu, which we can verify with both the CDC and Hopkins Medicine, claiming "because some of the symptoms of flu and COVID-19 are similar, it may actually be hard to tell the difference between them based on symptoms alone".[1]
- Variables with a strong correlation are Cold and Flu as well as Cold and Corona, which we can attribute to the fact that while the common cold shares symptoms with the flu, the symptoms of the flu are more intense, leading to less of a direct correlation.[2]

[1] https://www.cdc.gov/flu/symptoms/flu-vs-covid19.htm, https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/coronavirus-disease-2019-vs-the-flu
[2] https://www.cdc.gov/flu/symptoms/coldflu.htm

- Variables with not as strong of a correlation are Corona Silent and Cold, Flu, and Corona. We can attribute this to the ambiguity of corona silent as while someone may exhibit slight indication of cold or flu-like symptoms for a number of reasons, they could at the same time have a corona silent value > 0 simply because they are asymptomatic, creating a grey area as people start to get infected.[3]

Temporal Analysis

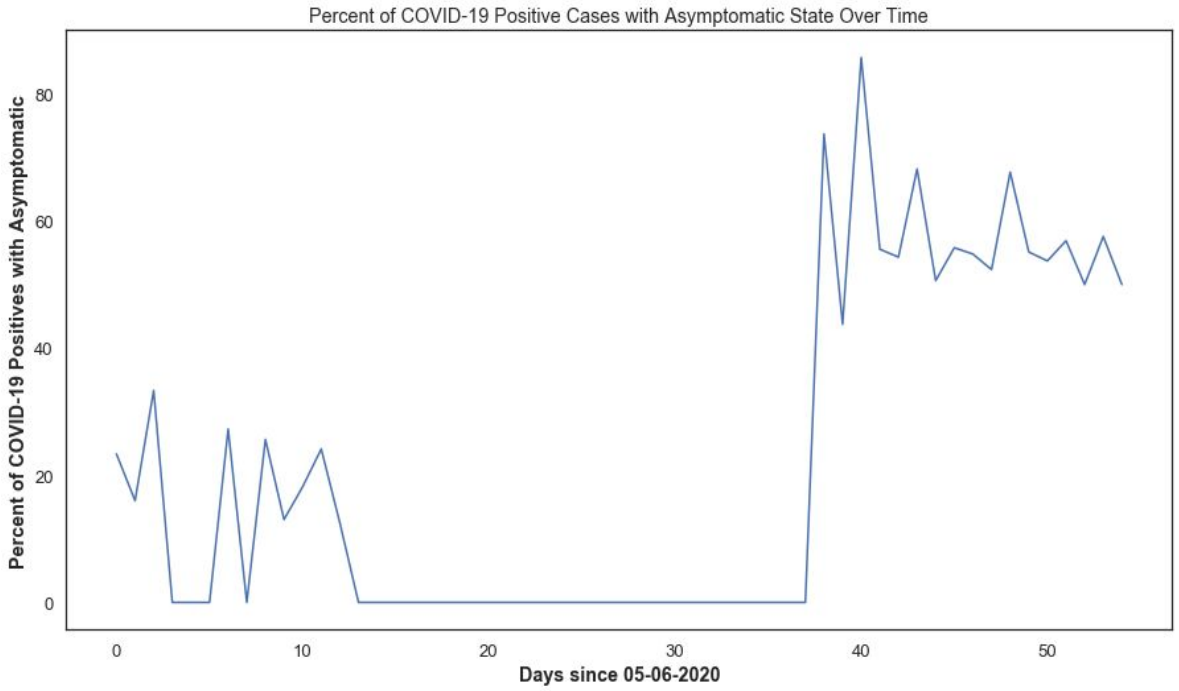
Applying the additional feature of the date for each record, we analyzed our discovered correlations in the context of over time. Here, we show a comparison between the total number of records each day, which indicates the amount of testing, and the number of records indicating an asymptomatic state with little symptoms each day. Because of the increase in testing in later time points, it can be observed that more cases of people who are asymptomatic are being caught with testing.



The number of asymptomatic cases each day can then be compared to the percent of COVID-19 positive cases that are asymptomatic, which can be interpreted as cases that are caught strictly from higher testing vs. cases from those with severe symptoms indicating COVID-19. As such, looking at these values helps understand the effectiveness of the testing and leads us to

[3] https://www.muhealth.org/our-stories/flu-cold-or-covid-19-consider-symptoms

prediction of any policy or societal changes involving testing frequency, something that has yet to be resolved.



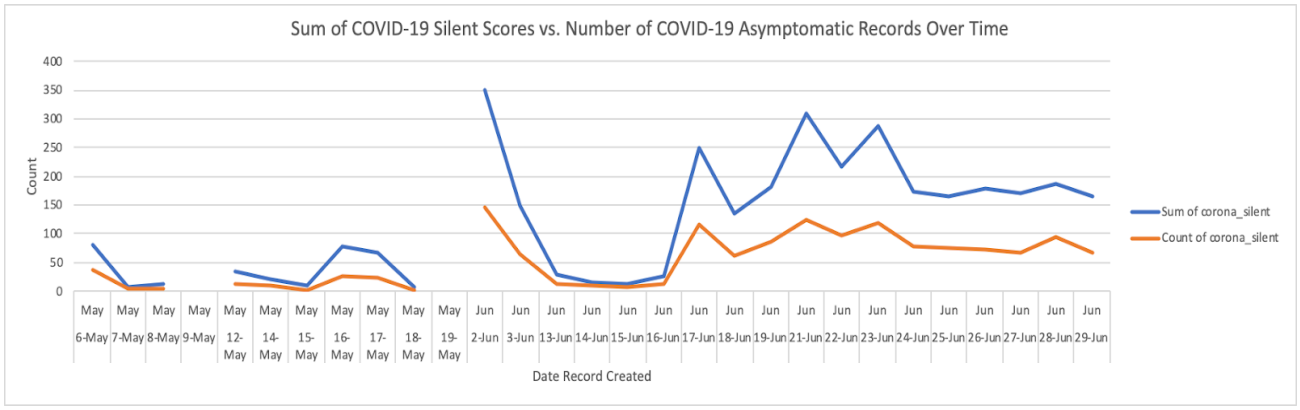
Temporal and Geographical Analysis

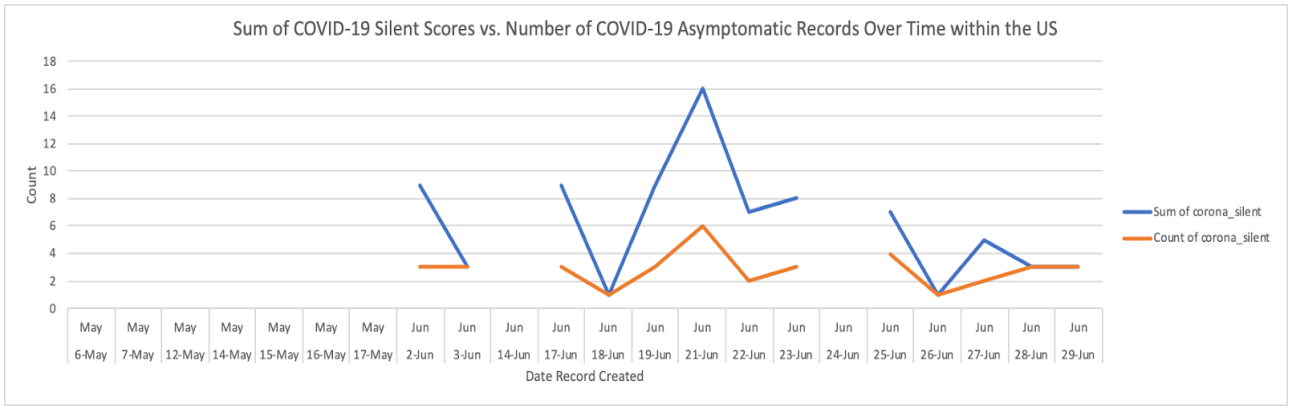
A step beyond analyzing testing frequency and positive cases over time is its application to a comparison between geographical areas. Much of the news content surrounding the pandemic involves how different countries are handling the pandemic differently, for instance:

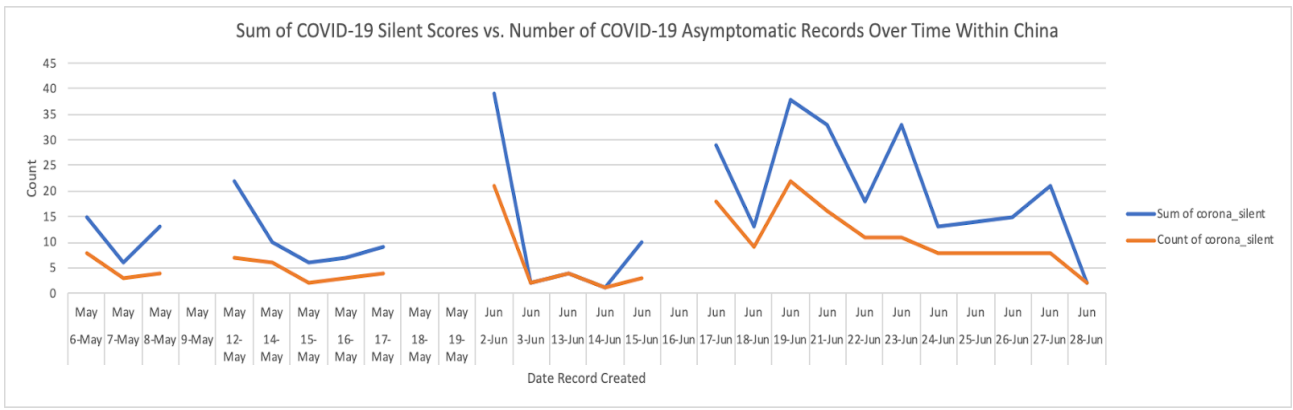
> "*US President Donald Trump says the US has done the "greatest testing" in the world - but Chinese state media says China has carried out three times more tests than the US."* - BBC World News[4]

From our analysis, we show a comparison of overall sum of Corona Silent scores vs. number of records over time and in specific countries, which can be interpreted as the severity of caught cases of COVID-19 when asymptomatic. Here, we show this analysis in all records and specifically within the U.S. and China with the time series plots below.

[4] https://www.bbc.com/news/world-us-canada-53221801







From these analyses, we determine while China's severity of cases of COVID-19 when asymptomatic generally decreases, that of the U.S. generally increases, which we can interpret as two ways:

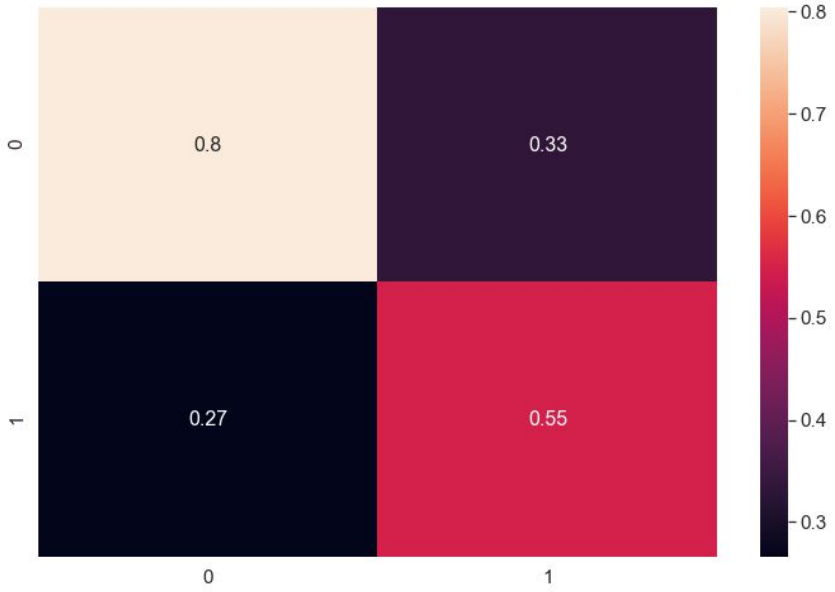
- <u>China has had more effective testing over time compared to the U.S., specifically for testing when people are asymptomatic.</u>
- The sample sizes within the dataset for these countries may be too small, so the analysis would have to be extended to the results of a greater number of recorded data.

Classification of COVID-19 PCR Test Result

Using our preliminary data analysis and machine learning, we created a model for classifying and predicting the result of a COVID-19 PCR test given values for each of the given features. The model we used is a CatBoost Regressor, an algorithm for gradient boosting on decision trees which has recently outperformed many existing boosting algorithms like XGBoost, Light GBM, etc. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of prediction models, typically decision trees, to effectively improve otherwise weaker models. CatBoost provides the additional incentive of using symmetric trees, allowing for a great decrease in training and prediction time especially for prediction of larger, high-scale datasets in production environments. Using this method, CatBoost followed the following procedure in its training:

1. Calculate residuals for each data point using a model that has been trained on all the **other data points at that time** (*For Example*, *to calculate residual for x5 datapoint, we train one model using x1, x2, x3, and x4*). Hence we train different models to calculate residuals for different data points. In the end, we are calculating residuals for each data point that the corresponding model has never seen that datapoint before.
2. Train the model by using the residuals of each data point as class labels
3. Repeat Step 1 & Step 2 (*for n iterations*)

From training CatBoost on all of our engineered features, we obtained an average accuracy (AUC) of 77%, ranging from 80% to 75% after Bayesian Hyperparameter Optimization. Below is an example of the confusion matrix generated from our model testing.



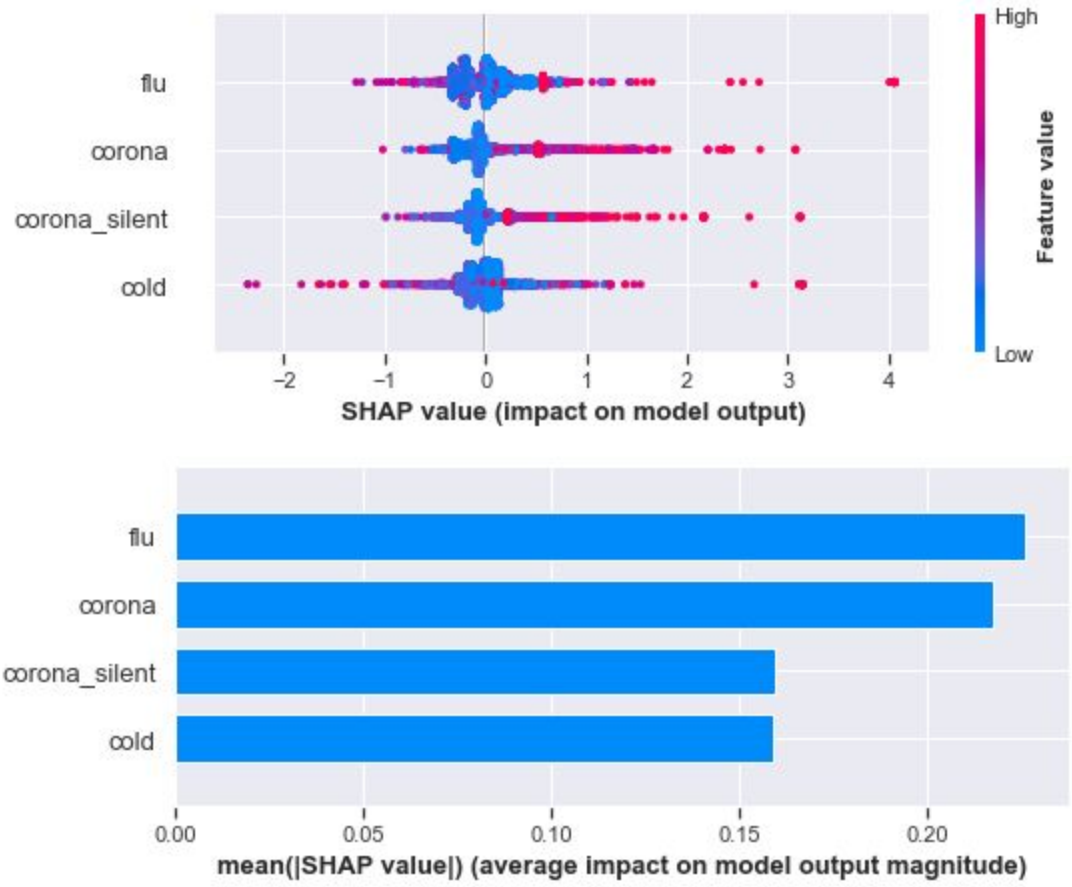
From this baseline model considering all symptom data, we applied the model for subsets of symptoms as well as standalone metrics and evaluated the accuracy on a holdout set of testing data. The results are organized in the table below.

Features used for training	**Model AUC**
All symptoms (Cold, Flu, Corona, Corona-Silent)	79%
Cold, Corona, Corona-Silent	75%
Corona, Corona-Silent	73%
Flu	61%
Cold	59%
Corona	59%
Corona-Silent	58%

From these results, we can verify with our correlation matrix which symptom vectors contribute most to the task of classifying for COVID-19 test results as we see the flu followed by the cold and corona have the highest standalone predictive accuracies. Further, we find specific combinations involving the cold, corona, and corona-silent are all very effective sets of features. These combinations are selected on the basis that because the flu and the common cold classify as different illnesses, removing them should ideally not only predict results for COVID-19, but also distinguish between COVID-19 and other viruses or illnesses moving forward.

Global Model Interpretation

From the trained CatBoost model, we apply model explainability frameworks LIME and SHAP for further understanding of the otherwise black box nature of machine learning models.

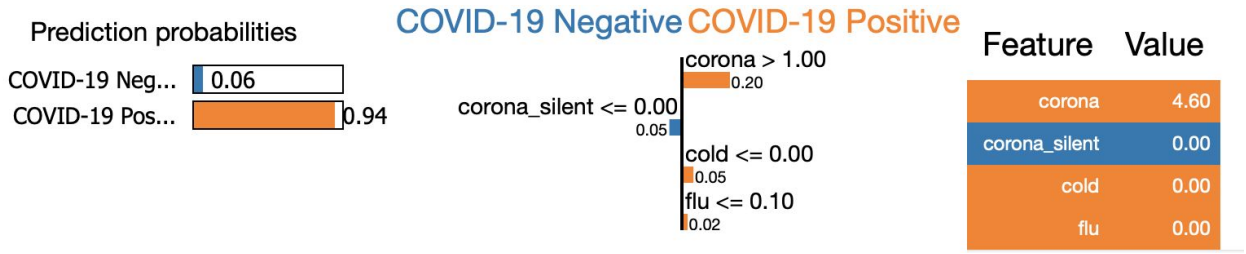


Local Model Interpretation

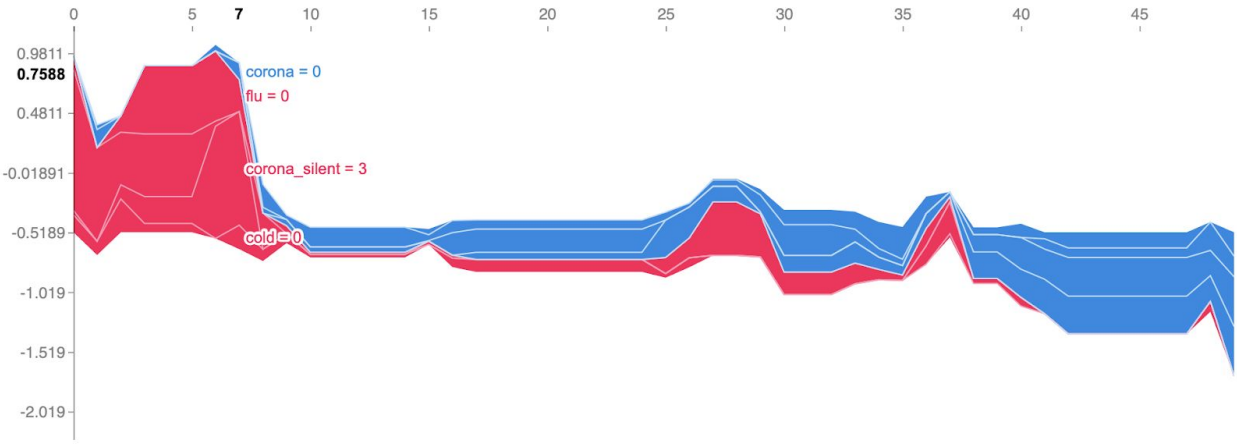
For local model interpretation, we added to our pipeline the ability to view the importance of each symptom in its prediction. Here, we use LIME for a visualization of the importance of each symptom and how it contributes to a positive or negative test result. LIME uses the following procedure:

1. Generates new samples then gets their predictions using the original model, and
2. Weighs these new samples by the proximity to the instance being explained.
3. Apply weights to interpret testing instances for feature importance.

Below is a sample of the figure generated by LIME for a specific test instance.



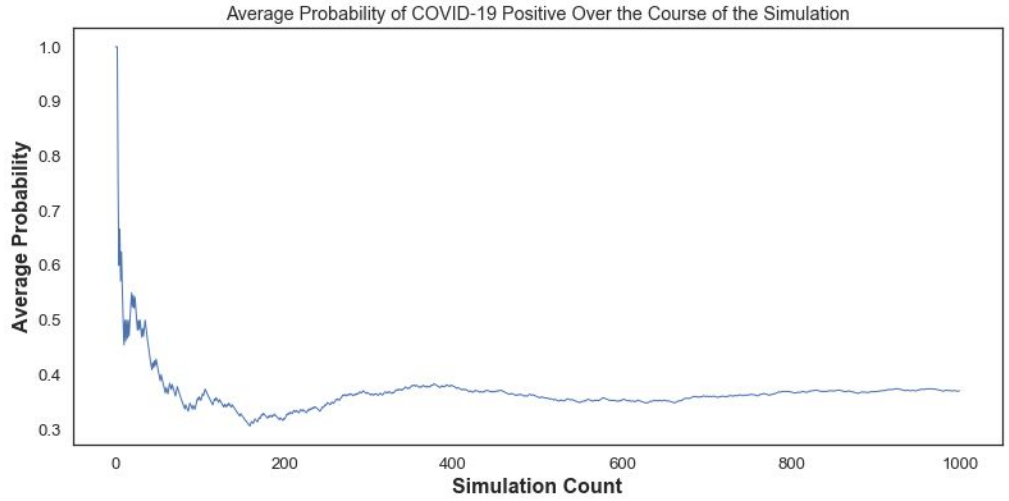
Applying this to a broader visualization, we used SHAP to generate a plot of the individual test instances and their feature importance over a variety of test cases. Below is a plot of each test instance on the x-axis and its respective feature importance values on the y-axis. These values can be viewed individually on the user interface and sorted based on individual features (e.g. temporal analysis by sorting by timepoints). Here, we group test cases by similarity, which can be observed with higher corona-silent values followed by higher cold and flu-like symptoms.

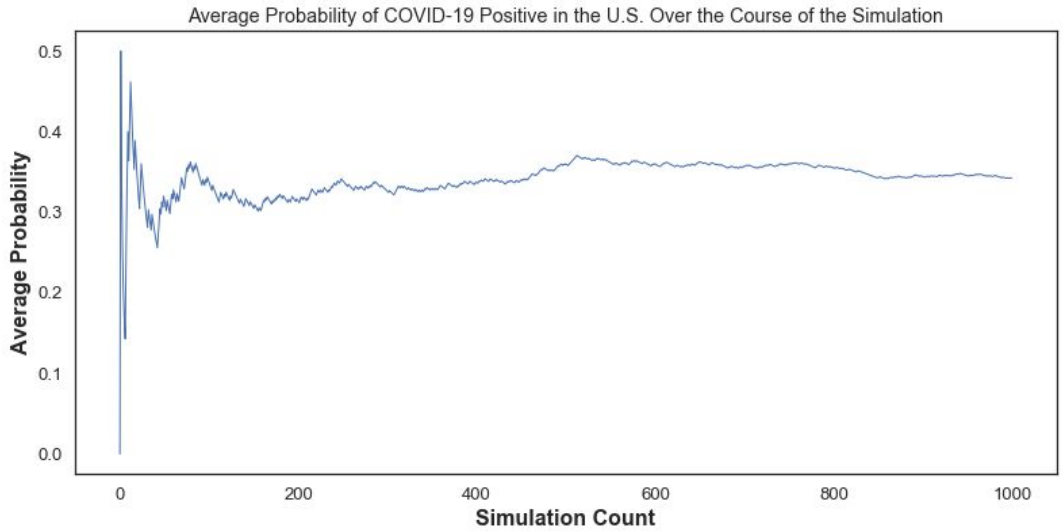


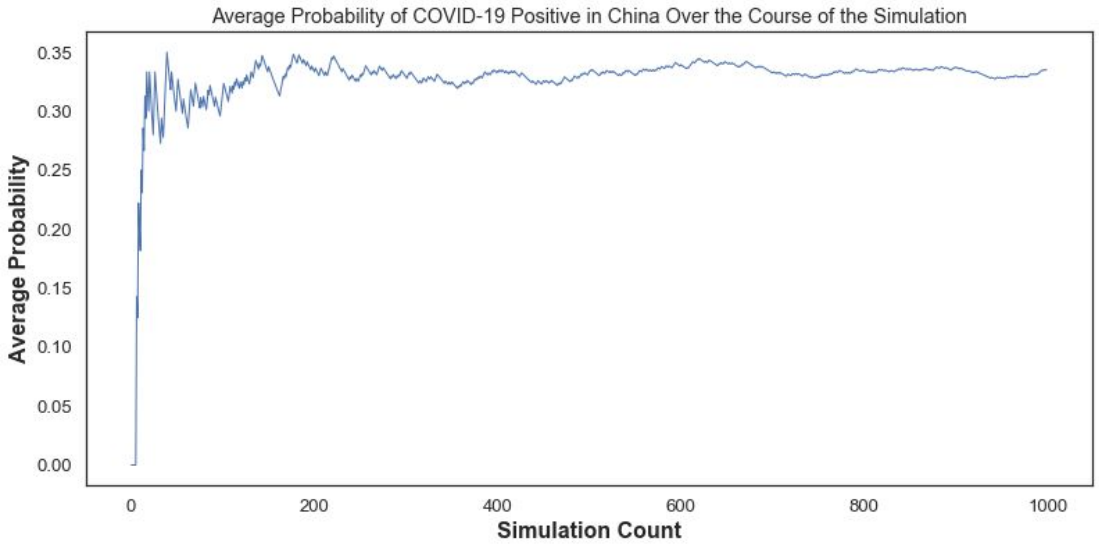
Probability Analysis

From the data provided, we also completed an analysis of the probabilities for specific samples of surveys to result in a positive test result, in effect predicting an "overall" COVID test result along with individual results. We do this using a Monte Carlo Simulation, which is a model used to predict the probability of different outcomes when the intervention of random variables is present. Monte Carlo simulations help to explain the impact of risk and uncertainty in prediction and forecasting models. Using this model, we generated plots for the average probability of an individual in a given population sample to test positive in the molecular diagnostic test for COVID-19. Because of the broad applicability of this simulation, we easily generated such plots

for both the entire dataset and for specific countries. Here, we show plots for overall and for both U.S. and China after 1000 runs of the simulation.

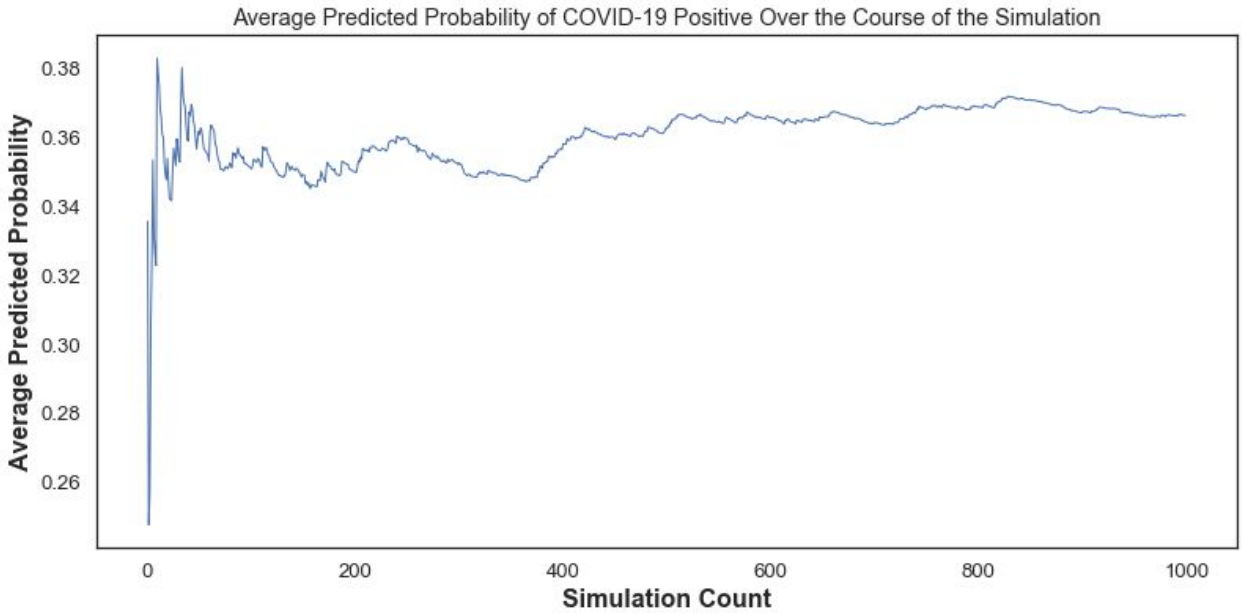






From these results, we can compare across geographical areas and, taking it one step further, time points, a COVID-19 probability density that provides information beyond defined statistical values.

Applying this to our machine learning model, we computed these same probabilities using the predicted COVID-19 Test Result:



Comparing the simulation results of our predictions with those of our given COVID-19 test results not only provides additional insight to any comparisons over geographical areas and/or time points, but it also further demonstrates the accuracy and efficacy of our classifier.

Conclusion

To summarize, we present in this report the following novel insights:

- Insight involving correlation between diagnostic vectors, date, and location both with test result and with other features, including a temporal and geographical analysis of specific features
- Prediction of COVID-19 PCR Test Result with up to 80% accuracy
- Interpretation of black box algorithm with plots of global feature importance and local feature importance for each prediction in testing dataset.

Using these insights, we hope future changes in science, society, and policies can reflect the ideas reflected here in our report.