Identification of Optimal Features for Effective COVID-19 Disease Forecasting with Machine Learning Models

Andy Chea [1]
andychea2000@gatech.edu

Saideep Narendrula [1]
snarendrula3@gatech.edu

Rachel Calder[1]
rachelcalder@gatech.edu

1. Department of Biosciences at the Georgia Institute of Technology in Atlanta, GA, USA

ABSTRACT

The start of the COVID-19 pandemic in the age of advancing data technology allowed for the curation of data at a quality and detail level that had not previously been observed. Much of this became open-source for the collective effort of reducing the harm of the pandemic. For example, Google released data on daily search trends per State for COVID-19-related symptoms [12]. Johns Hopkins University maintained databases of daily data on cases, deaths, and anonymized mobility data [13]. These readily available data sets provide researchers with the opportunity to build models to have ready for future communicable disease outbreaks. Black box machine learning (ML) models are trained for the forecasting of daily cases and deaths of COVID-19 from 2020-2022. Ablation procedures informed through the Pearson Correlation Coefficient showed that the number of features provided shows different results depending on the region of interest. The SHAP analysis provided insight on the usefulness of providing COVID related symptom search data to the ML models as it was the primary influence on the model's output. Utilizing ablations and SHAP, important features will be extracted to inform the next generation of these black box machine learning models.

CCS CONCEPTS

• Life and medical sciences • Machine Learning • Artificial intelligence

KEYWORDS

Explainable AI, Disease Forecasting, Cellular Mobility Data Tracking, Google Search Engine Trends, Xgboost, Linear Regression, SHAP

1 Introduction

Coined by Gunther Eysenbach in 2009 [5], "infodemiology" describes the new age of epidemiology using large swaths of open-source databases in tandem with data-mining approaches and novel data analysis to modernize the next insights in the field. This exciting era of Big Data is characterized by three "V" qualities: specific and disparate datasets (high variety), increasing numbers of samples and data points per sample (high volume), and rapid applicability and incorporation of these datums into machine learning models (high velocity) [8]. However, this final trait can sometimes clash with the first two. As the availability of new datasets balloons rapidly, how are data scientists to know which ones would be most imperative to add to their forecasters? In epidemiology, where decisive action at earlier time points can alter the entire trajectory of a pandemic, it is necessary to know which portions of the torrential incoming data stream are most informative and can be used to reduce suffering immediately.

The staggering abundance of available search engine data represents a veritable gold mine of human experience and curiosity, and this quantity of data compounds more and more with each day. In the wake of the COVID-19 pandemic that began in 2020, Google provided a means of understanding this new public health threat: users queried how and where the virus was spreading and how to best protect themselves and their families. While many studies have since emerged around the SARS-Cov2 virus and its epidemiology, we hope to use this period of quarantine and online pandemic interest to guide the construction of models for the next communicable disease. Even now in a post-vaccine world, COVID-19 is likely to remain an endemic threat that will need to be vigilantly contained with annual boosters and public awareness campaigns.

As symptoms precede diagnoses, it follows that an individual's recognition of their own symptoms and subsequent search could be used as a less intrusive means of predicting the arrival or resurgence of COVID-19 in a population. Alongside this public availability, search data often comes anonymized which means it can be examined without complex logistics or financial compensation. Survey data can provide more certain insights with targeted questions, however, surveys require active participation rather than the passive collection from search engines. Anonymized location data provides insight into compliance with social distancing policies, however the data collection is more intrusive. While many studies throughout the years have examined symptoms as early warning signals or the effectiveness of survey or location data, our project seeks to guide future work connecting the most useful data to devote resources to for rapid, actionable epidemiological insights by discovering informative variables.

1.1 Literature Review

When predicting waves of infection, epidemiologists can factor in mobility data and use the assumption that the length of time of proximity would increase the likelihood of new cases. However, mobility data should not be solely considered because it does not indicate other protective measures individuals could be using. For example, data from those in close proximity with preventative measures in place such as mask usage could cause an overestimate in the prediction of cases [7]. Mobility data can also vary based on

the demographics of the individuals surveyed, as the individuals must have access to and exposure to the technology used. Therefore, the data must be considered to be a sample of the total set.

The relationship between Google search trend data and COVID-19 cases has also been evaluated for predictive power. The assumption is that individuals may be more inclined to search for COVID-19 symptoms if they or someone they came into contact with are experiencing symptoms, and that search results could be used as a pseudo-diagnostic tool. A study using Google search trends in Turkey, Italy, Spain, France, and the United Kingdom found there to be significant time-varying correlations between pulmonary symptom search trends and new cases [3].

Current literature using search data to make epidemiological predictions has had mixed success. A pioneering paper from 2010 by Seifter et al [9] evaluated the effectiveness of the newly-released Google Trends database in mimicking seasonal Lyme disease trends. While rather simple twelve years later, this paper seemed to indicate then that search data was a useful resource in modeling. The authors also noted that the states where Lyme disease was considered endemic were also often the ones with the highest search quantity for terms like "Lyme disease" and "tick bites".

Then in 2013, Dugas et al [4] used the more specialized Google Flu Trends to train several forecasting models that predict the hospital load of flu patients in the following week. Their analysis was steeped in applicability as they focused only on data that was collectible in real-time and that could inform medical institutions at many levels. The most successful model was a generalized linear autoregressive moving average method (GARMA) with a negative binomial distribution. With a leave-one-out validation approach of 7 sets, they recorded an accuracy of 83%, lending credence to the validity of search data in our analysis. One caveat that they admit to is using only case data from one medical center, which might have overspecified the model. We will avoid this by examining cases from many different states to train a more generalizable model.

In 2017, a skeptical paper from Cervellin et al [2] claims that while previous work had found some correlation between search and epidemiological trends, there is much more nuance to consider before claiming these findings are substantial. Their experimental design found that both common diseases of low media coverage and rare, sensational diseases did not correlate well with their population prevalence. However, their methodology involves some very specific terms that even someone with the symptoms might not know to use when looking for more info, such as "renal colic" and "epistaxis." While all search terms are undoubtedly affected by media influence, these words might be too far removed from the average user to present a relevant test case. Our project utilizing symptom searches, which come more naturally than medical terminology, should prove to be informative.

In terms of prediction approaches, traditional machine learning methods have been used to compare against epidemiological models. In 2022, Rahman et al. [9] compared the accuracy of Xgboost compared to the ARIMA model with COVID-19 data in Bangladesh. In their approach, they found that ARIMA performed better than Xgboost in predicting cases and the number of deaths. With this approach, there was no explainability with their model so it was hard to see what features their Xgboost model was focusing on. There have been approaches to predicting cases and deaths of Covid 19 with ensemble models and explaining their approach with SHAP via Zheng et al [10] in 2022. Their approach utilized input features such as self-protection, policy indicators, community movement, and the time index. The SHAP toolbox they used provided information on the importance of the input features.

Also in 2022, CAMul (Calibrated and Accurate Multi-view Time-Series Forecasting) [6] has emerged as a new state-of-the-art multisource model that outperforms other leading forecasting models by 25% in accuracy and calibration metrics. It accounts for the relative relevance of different data views by assigning each a probabilistic uncertainty value and considering each in a context-specific manner, which are tasks often neglected in other multimodal models. CAMul applies to any multimodal question where an appropriate encoder can be used to transform the disparate data points into comparable values.

2 Proposal

2.1 Problem Statement

Simply put, our goal was to discover the input features that are most informative in training effective disease transmission machine learning models. We integrated state-level COVID-19 symptom search datasets with mobility and Facebook survey datasets with case and death data from each day in April 2020 and May 2021 to train different real-time models that predict COVID-19 incidence and mortality for the following week. To obtain a larger range of time, we considered data without the time range-limiting mobility data. This walk-forward approach will involve retraining the model while considering the data from each of the previous weeks. Extracting the most informative variables from the most successful models can then help inform other research groups about what types of data from our aggregated set are most vital to include in their forecasting tools.

2.2 Datasets

The Johns Hopkins University database of confirmed COVID-19 cases was used to capture daily cases and deaths within each American state from 2020 to 2021 and will be accessed through the Delphi Epidata API [13]. The Delphi Epidata API was also used to query data from the SafeGraph database based on daily metrics per state on how many patrons were visiting bars and restaurants. The SafeGraph database contains anonymized cell phone location data until April 19th, 2021 with flags for highly populated locations such as bars and restaurants. This date range shortened the range of comparable data from the other data sources, which extended to 2022. Therefore, two datasets were

generated for analysis, one with SafeGraph data and one without it but with a date range until May 2022. Also from the Delphi Epidata API, we gathered indicators of COVID-like illnesses from Facebook Survey data, using daily data with 7-day pooling. The Google COVID-19 Symptoms Search Trend database was used to gather counts of google searches potentially linked to COVID-19 symptoms [12].

Rhode Island and United States territories removed, as they were not consistently available across datasets. The datasets were merged according to state and date. For an exploratory analysis, we looked for days with missing data before they were aggregated into averages. This was where we discovered consistent missing data from 12-14-20 and 12-20-2020. It was common to view multi-day gaps along state data sources. These missing values were imputed with zeros before a 7-day smoothing occurred for all features except the pre-smoothed Facebook Survey data.

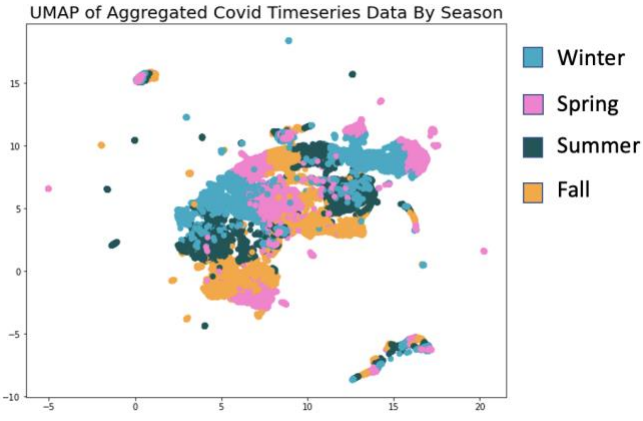


Figure 1: UMAP reduction of COVID Time-Series by Season

We produced a UMAP reduction analysis (Fig 1.) to explore relationships between weeks, months, and seasons of the year. We were pleased to see that there was variability in the data, indicating there are unique trends to observe. In Figure 1, we present the data colored by season. We found that clusters occurred by season and year, extending into Spring of 2022.

2.3 Methodology

Our approach utilizes machine learning models to evaluate the cases and deaths from COVID-19 in each state per week. For each week-length window, the data was smoothed using a 7-day average. Since the Facebook Survey data was provided with 7-day smoothing, it was not re-smoothed.

2.3.1 Response to Project Milestone Comments

We aimed to evaluate four models for time series forecasting: CAMul, Bidirectional LSTM, Xgboost, and SVM. After experimenting with CAMul, we chose not to use it as a forecaster to evaluate informative variables because it was not explainable by SHAP like the others. Bidirectional LSTM was another model we could not use due to the lack of compatibility with the current Tensorflow verison and SHAP. Since the model could not be explained, we decided to remove it from our list of forecasters. After considering feedback, we decided to not further pursue SVR as a forecasting method, Support Vector Machines (SVMs) are built upon drawing a line in hyperdimensional space that separates the different classes. Expanding this classification method into regression involves finding a hyperplane that includes a maximum number of training points. Forecasting using this method no longer seemed intuitive or practical because the temporal aspect changes each data point slightly in a way that might not necessarily be captured within our input features. While this fact has not excluded it from being used in the past as an epidemiological predictor, the fact that it is also not explainable with SHAP makes it less attractive for our analysis.

Xgboost (XGB), Random Forests, and Linear Regression were used as the machine learning models.

In its place as one of the simpler models, we opted for a Random Forest Regressor (RFR). This ML model is an ensemble that averages the outputs of various bagged decision trees to ultimately decide the value of the next prediction. Every tree is given a unique bagged dataset that is generated from the input data using sampling with replacement. Additionally, each branch within a tree is only allowed to be selected from a random subset of the features, which leads to a large degree of individuality in how each decision tree arrives at its prediction. While the underlying algorithm is still viewed more as a classifier, we decided that the inclusion of an ensemble model might broaden the model space we were examining. In order to train this model, missing numeric data had to be imputed using a backfill method to pull the most recent value of the time series into the blank spaces. XGBoost was selected as a representative gradient boosting algorithm. Xgboost is commonly used in research to understand the variable importance of a feature dense dataset. Our choice in Xgboost as a regressor was primarily to understand the importance of features and not to improve the predictive power.

The models can be implemented with the Xgboost and sci-kit learn packages in Python. Initial results performed well in approximating the ground truth models, so the marginal gain from implementing optimization of the hyperparameters was deemed unnecessary.

To train each model, we used a train-validation approach. The 2020-2021 data will be used to train the model. To validate each model, we used a walk-forward validation approach every week in the 2022 data and averaged the results from each walk.

To evaluate our models' efficacy, we used two primary metrics: MAE (Mean Absolute Error) and RMSE (Root-Mean-Square Error). MAE is the simplest metric and represents the sum of absolute errors divided by the sample size. This metric is intuitive, and better models will have lower values.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

RMSE is calculated similarly as a function of the difference between the expected and observed values but is more responsive to outliers due to the absolute errors being squared prior to averaging. Large differences thus incur more error than in MAE. In our experiment where we want to produce the best forecaster, disincentivizing large differences makes this popular metric crucial to consider.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}$$

We used two methods to explain our model outputs: ablation and SHAP values.

One method we will use to explain our model is to conduct ablations on each model. To get an important feature list, we will use the Pearson Correlation Coefficient on each feature in the dataset. We will order the feature list by the resulting coefficient. Using this list, we will train Xgboost with 25 features at a time. This process will be repeated for our interested states. We will evaluate model performance through the metrics described early. We will plot model performance against the number of features that the model has.

To explain the model embeddings, our team computed SHAP (SHapley Additive exPlanation) [8] values to get individual feature importance. The SHAP value explains how each feature changes the model prediction. The SHAP values provide explanations of model embeddings with no information on input features. For the tree-based model, Xgboost, the SHAP values can be computed using local model summarizations [9].

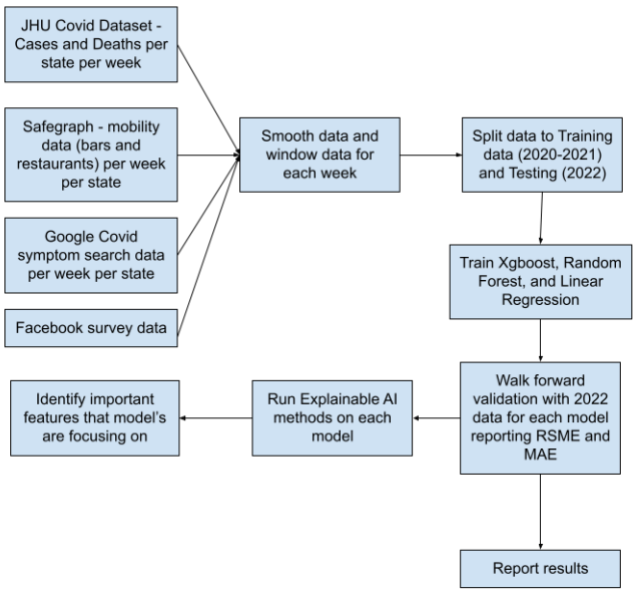


Fig 2. Flow chart of projected milestone steps.

3 Results and Discussion

3.1 Experiments

With our experiment, we used ablations to answer 1. How does model performance change with the addition of less important features? 2. Can we see if the PCC value was impacting the model performance?

Our SHAP results will answer 3. Which expected and unexpected features show up most frequently as informative in generating accurate predictions?

3.2 Results

3.2.1 Forecasting Performance

Overall, the values of MAE and RMSE were quite low due to applying the standard scaler on each column- only linear regression had any result values above 1, indicating that it performed worse in error reduction than either of the machine learning models.

Focusing on MAE, we note an overall lower error across the five states in the cases result than the deaths result, with the exception of California's linear regression result. While both XGBoost and Random Forest had similar MAE's below 0.05, linear regression performed significantly worse with an average MAE of 0.909 across the five states. This difference could be due to the fact that linear regression cannot generalize well to temporal data. While some of the effects of timing are contained in the trends of the features, the loss in ordering from past to future might amplify the error.

In all states and all models, the RMSE was always higher than MAE, meaning that there were large differentials in the prediction and ground truth that created bigger error values after being squared. Across all models, RMSE of case forecasting was lower than RMSE of death forecasting all of the states except one: California. It seems therefore that cases are marginally easier to forecast than deaths in many regions. Compared to the other analyzed states, California might have a dense enough population to lead to rapid changes in deaths that are more difficult to forecast.

This reveals that in terms of best performance for both metrics, both the machine learning models performed better than the linear regression, and that there were large differences in some of the walk-forward validation predictions that were emphasized more in the calculation of RMSE than in MAE.

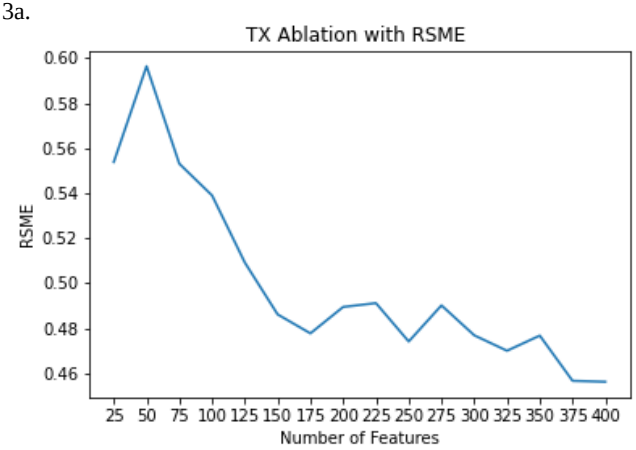
3.2.2 Ablations

From our ablation experiments on each state, we found two interesting observations. Our ablation procedure explained in the methods, was performed on the Texas data Fig 3a. The RSME gradually decreased as the number of features increased at 25 per iteration. This response is expected as the number of features

should lead to better performance in the prediction power of the model. The starting RSME was much higher than we expected however we attributed that to the high collinearity features that were fed into the model. GA, MA, and CA all performed similarly to TX. Our ablation procedure performed on the Florida dataset showed that increasing the features did not result in a lower RSME Fig 3b. There was a spike in the resulting RSME at half of the features provided. The rise and decrease of the resulting RSME as the features increased led to the same starting and ending RSME.

The first observation is easily explained as introducing more data led to a better performance. The features were ordered from positive to negative correlation so the introduction of the negatively correlation features had a positive impact on model performance. However the observation on the FL dataset was different. As less positive correlated features were provided the model performed worse and once negatively correlated features were fed in, the model performance was better. The SHAP beeswarm plot for FL shows that three out of the top five features were negatively correlated with the resulting prediction feature, the number of deaths in the state for that day. Considering the type of data that is gathered during a pandemic, gathering relevant search query data is more important.

3a.



3b.

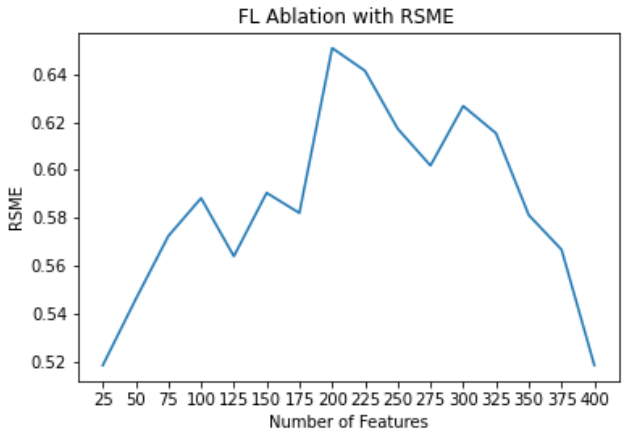


Figure 3: **XGboost Model Performance with Ablation procedure for Texas (3a) and Florida (3B).** Texas and other states show expected results with model performance increasing as the number of features increased. However, Florida showed an interesting decrease of performance as the number of features increased.

3.2.3 SHAP Results

Examining the SHAP values for different models reveals different informative variables for each of the queried states. The representative beeswarm plots display the top 9 most informative features the trained model uses in generating its next forecast from most to least. By default, the algorithm weights features of average consistent information more highly than features that are highly informative in only a subset of samples.

XGBoost's top feature was facebook survey data in 7 of the 10 walk-forward validations, and it appeared in the top 9 in all of them. Several symptoms appear in more than one state's top nine features in predicting either cases or deaths for the next day. Among the most common are croup, acute bronchitis, aphonia (voice loss), and cough. These are all symptoms that have an intuitive connection to the respiratory virus SARS-CoV2. Croup appears in half of the XGB important feature sets and is a barking cough that signals an infection of the upper respiratory system. The latter three are all reasonable symptom searches preceding a COVID diagnosis. Some surprisingly informative symptoms implicated among the top 9's were gingivitis, dementia, and stroke, but could indicate how older populations more at risk for COVID were searching for health issues. Comorbidity is another explanation rather than a direct causal relationship between the implicated signal symptom and COVID response, where the same underlying mechanisms manifest themselves as both vulnerability to COVID and the given symptom in the searcher. This indirect relationship could help explain some of the least expected symptom appearances, such as alcoholism or ear pain.

Random Forest had a much more focused set of explanatory symptoms, with a smaller set of features appearing repeatedly. The most frequent symptoms showed up in four of the ten runs, and among them are some that are seemingly not related to COVID: astigmatism and motion sickness. These do not necessarily correlate with older populations, so the earlier possible indirect link does not seem applicable. However, many of the intuitive symptoms from the XGB set also appear more frequently, such as nasal congestion, laryngitis, and sore throat. Cough, croup, and upper respiratory tract infection show up at similar frequencies.

Linear regression homes in even more on canonically expected COVID symptoms, with nasal congestion and cough becoming top features in over half of the forecasts done. Respiratory symptoms like pneumonia and common cold als Unlike the previous two models,, it does not have the Facebook survey data in its top nine features for every forecast. In all, it seems that linear regression uses other respiratory symptoms better as signals in its predictions than the two ML models, but has more error than

Looking broadly at the feature overlap between the models, we can extract that due to linear regression's stronger focus on typical

symptoms associated with respiratory diseases reduces the overlap of the three symptom sets to that. While it is reassuring that the models placed great emphasis on the same factors we think are important for predicting COVID spread and mortality, it does not reveal many novel features that were present in a lot of the forecasting decisions that could then be applied in future training sets. However, our work shores up which of the symptom search features currently collected are the most imperative to include when fitting a model.

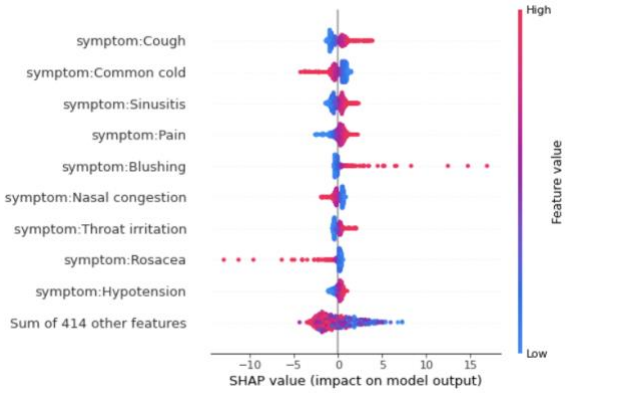


Figure 4: **California Linear Regression SHAP Beeswarm: Majority "Intuitive" COVID Predictors**: This SHAP was chosen from the total set of 30 to represent a top 9 feature set with many symptom searches that could relate to respiratory viruses like COVID.

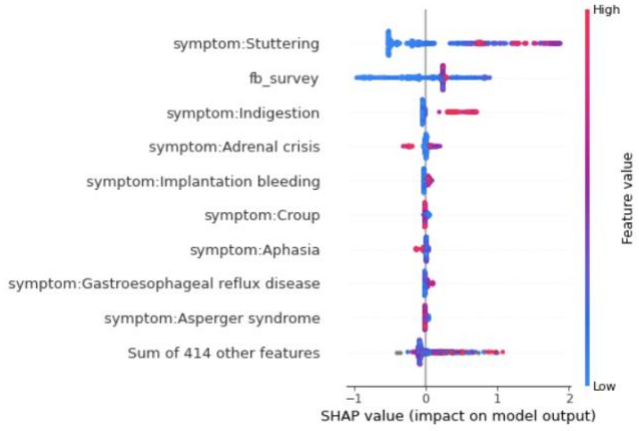


Figure 5: **Massachusetts XGBoost SHAP Beeswarm: Less "Intuitive" COVID Predictors:** Symptom search data on adrenal crisis, stuttering, and aphasia are not readily linkable to a respiratory disease like COVID-19 at this time.

3.3 Discussion

Since our work seeks to inform future forecasting models, we consider in this discussion the logistics of the different models used. Temporally, performing the fitting in a walk-forward validation took the most amount of time for the RFR, as building an ensemble of decision trees is costly. In order to mimic the widescale deployment methods that could be built from our suggested features, we sought out the simplest implementations for each of the algorithms, and fortunately each is available easily through Python packages. The more complex models that were not SHAP compatible also come with a steeper learning curve that may be less practical for those less-versed in analytics. In general, for these simulations of real-time forecasting, the simpler models often take less time to incorporate new data and tune hyperparameters, and can often generate solid ballpark approximations, as shown by the low errors detected in our results. These are huge ramifications for applicability during a pandemic.

A factor that might have affected our data is the degree of imputation of the different signals. Within the 400+ symptom search daily counts, each symptom had a different proportion of blanks and later NaNs in our numerical datasets. For the algorithms that are not tolerant to those, our choice to impute data means that some of the training data is simply back-filled from the last day that had an observation. This might differentially advantage different features depending on the trends in the response at that point in the time series.

In conclusion, self-reported data continues to be a descriptive and easily-accessible wealth for forecasting. We showed that of the several possible symptoms of the various strains of COVID, that certain symptoms like "cough" and "congestion" are indispensable in training. Our work also implicates some surprising search data that informed accurate COVID trends, such as astigmatism and motion sickness that might be implemented should the data also be available. Future work should include combing more models for informative signals through finding novel ways to harmonize SHAP with them. The repository of unexpected signals should continue to be grown so that forecasters can get the best marginal gain from addition of new features. The more work that is done to elucidate the explanatory features, the more efficiently modelers can rapidly select and implement new data as we prepare for future epidemics.

REFERENCES

[1] James Bergstra, Daniel Yamins, and David Cox. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In Proceedings of the 30th International Conference on Machine Learning, PMLR, 115–123. Retrieved July 17, 2022 from https://proceedings.mlr.press/v28/bergstra13.html

[2] Gianfranco Cervellin, Ivan Comelli, and Giuseppe Lippi. 2017. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. Journal of Epidemiology and Global Health 7, 3 (September 2017), 185–189. DOI:https://doi.org/10.1016/j.jegh.2017.06.001

[3] Halit Cinarka, Mehmet Atilla Uysal, Atilla Cifter, Elif Yelda Niksarlioglu, and Aslı Çarkoğlu. 2021. The relationship between Google search interest for pulmonary symptoms and COVID-19 cases using dynamic conditional correlation analysis. Sci Rep 11, 1 (July 2021), 14387. DOI:https://doi.org/10.1038/s41598-021-93836-y

[4] Andrea Freyer Dugas, Mehdi Jalalpour, Yulia Gel, Scott Levin, Fred Torcaso, Takeru Igusa, and Richard E. Rothman. 2013. Influenza Forecasting with Google Flu Trends. PLOS ONE 8, 2 (February 2013), e56176. DOI:https://doi.org/10.1371/journal.pone.0056176

[5] Gunther Eysenbach. 2009. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. Journal of Medical Internet Research 11, 1 (March 2009), e1157. DOI:https://doi.org/10.2196/jmir.1157

[6] Harshavardhan Kamarthi, Lingkai Kong, Alexander Rodríguez, Chao Zhang, and B. Aditya Prakash. 2022. CAMul: Calibrated and Accurate Multi-view Time-Series Forecasting. DOI:https://doi.org/10.48550/arXiv.2109.07438

[7] Nishant Kishore. 2021. Mobility data as a proxy for epidemic measures. Nat Comput Sci 1, 9 (September 2021), 567–568. DOI:https://doi.org/10.1038/s43588-021-00127-7

[8] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. Retrieved November 4, 2022 from http://arxiv.org/abs/1705.07874

[9] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2, 1 (January 2020), 56–67. DOI:https://doi.org/10.1038/s42256-019-0138-9

[10] Stephen J Mooney, Daniel J Westreich, and Abdulrahman M El-Sayed. 2015. Epidemiology in the Era of Big Data. Epidemiology 26, 3 (May 2015), 390–394. DOI:https://doi.org/10.1097/EDE.0000000000000274

[11] Ari Seifter, Alison Rebman, Kate Geis, and John Aucott. 2010. The utility of “Google Trends” for epidemiological research: Lyme disease as an example. Geospatial health 4, (May 2010), 135–7. DOI:https://doi.org/10.4081/gh.2010.195

[12] Hu-Li Zheng, Shu-Yi An, Bao-Jun Qiao, Peng Guan, De-Sheng Huang, and Wei Wu. 2022. A data-driven interpretable ensemble framework based on tree models for forecasting the occurrence of COVID-19 in the USA. Environ Sci Pollut Res (September 2022). DOI:https://doi.org/10.1007/s11356-022-23132-3

[13] Xun Zheng, Manzil Zaheer, Amr Ahmed, Yuan Wang, Eric P. Xing, and Alexander J. Smola. 2017. State Space LSTM Models with Particle MCMC Inference. Retrieved November 4, 2022 from http://arxiv.org/abs/1711.11179

[14] Explore COVID-19 Symptoms Search Trends. Retrieved October 6, 2022 from https://pair-code.github.io/covid19_symptom_dataset

[15] Epidata API Home. Delphi Epidata API. Retrieved October 6, 2022 from https://cmu-delphi.github.io/delphi-epidata/

Identification of Optimal Features for Effective COVID-19 Disease Forecasting with Machine Learning Models

Andy Chea [1]
andychea2000@gatech.edu

Saideep Narendrula [1]
snarendrula3@gatech.edu

Rachel Calder[1]
rachelcalder@gatech.edu

1. Department of Biosciences at the Georgia Institute of Technology in Atlanta, GA, USA

ABSTRACT

The start of the COVID-19 pandemic in the age of advancing data technology allowed for the curation of data at a quality and detail level that had not previously been observed. Much of this became open-source for the collective effort of reducing the harm of the pandemic. For example, Google released data on weekly search trends per State for COVID-19-related symptoms [12]. Johns Hopkins University maintained databases of weekly data on cases, deaths, and anonymized mobility data [13]. These readily available data sets provide researchers with the opportunity to build models to have ready for future communicable disease outbreaks. Black box machine learning models are trained for the forecasting of weekly cases and deaths of COVID-19 from 2020-2022. Utilizing ablations and SHAP, important features will be extracted to evaluate the efficacy of these black box machine learning models.

CCS CONCEPTS

• Life and medical sciences • Machine Learning • Artificial intelligence

KEYWORDS

Explainable AI, Disease Forecasting, Cellular Mobility Data Tracking, Google Search Engine Trends, Xgboost, SVM, Linear Regression, SHAP

1 Introduction

Coined by Gunther Eysenbach in 2009 [5], "infodemiology" describes the new age of epidemiology using large swaths of open-source databases in tandem with data-mining approaches and novel data analysis to modernize the next insights in the field. This exciting era of Big Data is characterized by three "V" qualities: specific and disparate datasets (high variety), increasing numbers of samples and data points per sample (high volume), and rapid applicability and incorporation of these datums into machine learning models (high velocity) [8]. However, this final trait can sometimes clash with the first two. As the availability of new datasets balloons rapidly, how are data scientists supposed to know which ones would be most imperative to add to their forecasters? In epidemiology, where decisive action at earlier time points can alter the entire trajectory of a pandemic, it is necessary to know which portions of the torrential incoming data stream are most informative and can be used to reduce suffering immediately.

The staggering abundance of available search data represents a veritable gold mine of human experience and curiosity, and this quantity of data compounds more and more with each day. In the wake of the COVID-19 pandemic that began in 2020, Google provided a means of understanding this new public health threat: users queried how and where the virus was spreading and how to best protect themselves and their families. While many studies have since emerged around the SARS-Cov2 virus and its epidemiology, we hope to use this period of quarantine and online pandemic interest to guide the construction of models for the next communicable disease. Even now in a post-vaccine world, COVID-19 is likely to remain an endemic threat that will need to be vigilantly contained with annual boosters and public awareness campaigns.

As symptoms precede diagnoses, it follows that an individual's recognition of their own symptoms and subsequent search could be used as a less intrusive means of predicting the arrival or resurgence of COVID-19 in a population. Alongside this public availability, search data often comes anonymized which means it can be examined without complex logistics or financial compensation. While many studies throughout the years have examined symptoms as early warning signals, our project seeks to guide future work connecting symptom search data to rapid, actionable epidemiological insights by discovering informative variables.

1.1 Literature Review

When predicting waves of infection, epidemiologists can factor in mobility data and use the assumption that the length of time of proximity would increase the likelihood of new cases. However, mobility data should not be solely considered because it does not indicate other protective measures individuals could be using. For example, data from those in close proximity with preventative measures in

place such as mask usage could cause an overestimate in the prediction of cases [7]. Mobility data can also vary based on the demographics of the individuals surveyed, as the individuals must have access to and exposure to the technology used. Therefore, the data must be considered to be a sample of the total set.

The relationship between Google search trend data and COVID-19 cases has also been evaluated for predictive power. The assumption is that individuals may be more inclined to search for COVID-19 symptoms if they or someone they came into contact with are experiencing symptoms, and that search results could be used as a pseudo-diagnostic tool. A study using Google search trends in Turkey, Italy, Spain, France, and the United Kingdom found there to be significant time-varying correlations between pulmonary symptom search trends and new cases [3].

Current literature using search data to make epidemiological predictions has had mixed success. A pioneering paper from 2010 by Seifter et al [9] evaluated the effectiveness of the newly-released Google Trends database in mimicking seasonal Lyme disease trends. While rather simple twelve years later, this paper seemed to indicate then that search data was a useful resource in modeling. The authors also noted that the states where Lyme disease was considered endemic were also often the ones with the highest search quantity for terms like "Lyme disease" and "tick bites".

Then in 2013, Dugas et al [4] used the more specialized Google Flu Trends to train several forecasting models that predict the hospital load of flu patients in the following week. Their analysis was steeped in applicability as they focused only on data that was collectible in real-time and that could inform medical institutions at many levels. The most successful model was a generalized linear autoregressive moving average method (GARMA) with a negative binomial distribution. With a leave-one-out validation approach of 7 sets, they recorded an accuracy of 83%, lending credence to the validity of search data in our analysis. One caveat that they admit to is using only case data from one medical center, which might have overspecified the model. We will avoid this by examining cases from many different states to train a more generalizable model.

In 2017, a skeptical paper from Cervellin et al [2] claims that while previous work had found some correlation between search and epidemiological trends, there is much more nuance to consider before claiming these findings are substantial. Their experimental design found that both common diseases of low media coverage and rare, sensational diseases did not correlate well with their population prevalence. However, their methodology involves some very specific terms that even someone with the symptoms might not know to use when looking for more info, such as "renal colic" and "epistaxis." While all search terms are undoubtedly affected by media influence, these words might be too far removed from the average user to present a relevant test case. Our project utilizing symptom searches, which come more naturally than medical terminology, should prove to be informative.

In terms of prediction approaches, traditional machine learning methods have been used to compare against epidemiological models. In 2022, Rahman et al. [9] compared the accuracy of Xgboost compared to the ARIMA model with COVID-19 data in Bangladesh. In their approach, they found that ARIMA performed better than Xgboost in predicting cases and the number of deaths. With this approach, there was no explainability with their model so it was hard to see what features their Xgboost model was focusing on. There have been approaches to predicting cases and deaths of Covid 19 with ensemble models and explaining their approach with SHAP via Zheng et al [10] in 2022. Their approach utilized input features such as self-protection, policy indicators, community movement, and the time index. The SHAP toolbox they used provided information on the importance of the input features.

Also in 2022, CAMul (Calibrated and Accurate Multi-view Time-Series Forecasting) [6] has emerged as a new state-of-the-art multisource model that outperforms other leading forecasting models by 25% in accuracy and calibration metrics. It accounts for the relative relevance of different data views by assigning each a probabilistic uncertainty value and considering each in a context-specific manner, which are tasks often neglected in other multimodal models. CAMul applies to any multimodal question where an appropriate encoder can be used to transform the disparate data points into comparable values.

1.3 Feedback Improvements

In expanding our model evaluation time, we will also explore CAMul as a predictive model and use its attention weights to implicate the most informative variables prior to using SHAP on the intermediate outputs of each submodel. To create a more realistic real-time model, we will heed the recommendation and switch from a full-year prediction to a weekly train-test cycle that can incorporate as much of the available data for a given time point as possible

While we did consider the inclusion of saliency maps as an additional explainable AI technique, an appropriate methodology to harmonize it with the model outputs could not be identified. Our understanding that saliency maps are used primarily in computer vision to determine the most informative regions of a picture did not mesh well with our non-image outputs. Instead, we will still use both ablation and SHAP values to explain the feature embeddings of the four models. We have elaborated further on our SHAP procedure which should be adequately informative.

Lastly, we will explore the lead times of the different models at different temporal (1-, 2-, and 4-week) and geographical (national vs state) scales. This will broaden our work to not only describe which data might be the most important in epidemic forecasting for each model, but also the ideal efficacy of the earliest warning. The number of different subsets explored for lead time will be dependent on the remaining time following completion of the explainable AI methodology.

2 Proposal

2.1 Problem Statement

Simply put, our goal is to discover the input features that are most informative in training effective disease transmission machine learning models. We will integrate state-level COVID-19 symptom search datasets with mobility and case datasets from each week in 2020 and 2021 to train different real-time models that predict COVID-19 incidence and mortality for the following week. Since the outset, we have also added a Facebook survey of COVID-like symptoms. This walk-forward approach will involve retraining the model while considering the data from each of the previous weeks. Extracting the most informative variables from the most successful models can then help inform other research groups about what types of data from our aggregated set are most vital to include in their forecasting tools.

2.2 Datasets

The Johns Hopkins University database of confirmed COVID-19 cases will be used to capture weekly cases and deaths within each American state from 2020 to 2021 and will be accessed through the Delphi Epidata API [13]. The Delphi Epidata API was also used to query data from the SafeGraph database based on weekly metrics per state on how many patrons were visiting bars and restaurants. The SafeGraph database contains anonymized cell phone location data until April 19th, 2021 with flags for highly populated locations such as bars and restaurants, which limited the range of comparable data from the other data sources, which extended to 2022. We are considering removing the SafeGraph data if we find we need a larger range of data. Also from the Delphi Epidata API, we gathered indicators of influenza-like illnesses from Facebook Survey data, which was an addition from our previous proposal. The Google COVID-19 Symptoms Search Trend database will be used to identify counts of google searches mapped back to COVID-19 symptoms [12].

All datasets were taken as raw day values for the given date range. Then, they were averaged by state by week. Rhode Island and Puerto Rico were removed, as they were not consistently available across datasets. The datasets were merged according to state and date. There was consistently missing data between 12-14-20 and 12-20-2020. To resolve this, we took the average of the next and prior weeks as the value.

For an exploratory analysis, we looked for days with missing data before they were aggregated into averages. This was where we discovered the consistent missing data from 12-14-20 and 12-20-2020. There were several states which lacked data week to week.

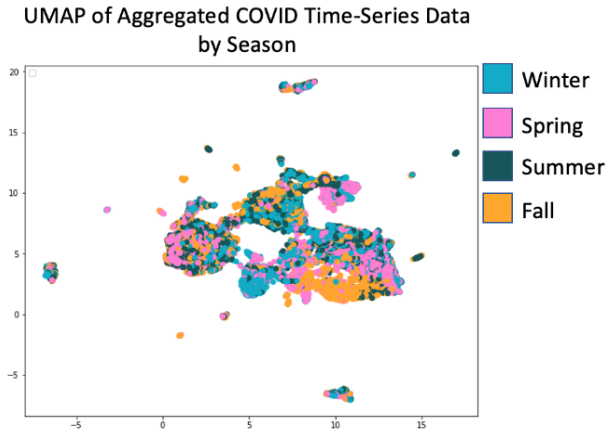


Fig 1. UMAP reduction of COVID Time-Series by Season

We produced a UMAP reduction analysis (Fig 1.) to explore relationships between weeks, months, and seasons of the year. We were pleased to see that there was variability in the data, indicating there are unique trends to observe. In Figure 1, we present the data colored by season. We are interested in exploring these trends further to visualize weeks with peak cases or that align with flu season.

2.3 Methodology

Our approach will utilize machine learning models to evaluate the cases and deaths from Covid-19 in each state per week. The mobility and symptom data will be

condensed into weekly data to match the JHU weekly reporting of the deaths and cases. For each week-length window, the data will be smoothed using Locally Weighted Scatterplot Smoothing (LOWESS) [10] and statistics will be computed on each window. Since the JHU case data is already reported at the weekly level, we will utilize LOWESS to smooth the data.

We will evaluate four models for time series forecasting: CAMul, Bidirectional LSTM, Xgboost, and SVM.

CAMul [6] has four major sections to the model: multi-view probabilistic encoders, view specific correlation graph, a context-specific dynamic view selection, and a forecasting distribution decoder. The multi-view probabilistic encoder learns the data embedding for each data point. The view-specific correlation graph encodes the stochastic relationship between the input data and reference points. The context-specific dynamic view selection integrates the embeddings using importance weights from a cross-attention module. The forecasting distribution decoder learns the probabilistic output distribution from the attention module embedding.

Long-Term Short Memory (LSTM) [11] networks are useful for time-based sequence learning. The RNN focuses on defining the distribution of each observation based on the past through a neural network. This network is commonly used for forecasting models as it maintains temporal information when a bi-directional LSTM module is used. The embedding of the Bi-LSTM modules should provide information on how each feature is affecting the model's output.

Xgboost and SVM will be used as the machine learning models. Each model can be implemented with the Xgboost and sci-kit learn packages. To optimize the model's hyperparameters, Bayesian Optimization will be performed to identify the best set of parameters utilizing hyper-opt [1].

To train each model, we will use a train-validation approach. The 2020-2021 data will be used to train the model. To validate each model, we will use a walk-forward validation approach every week in the 2022 data and average the results from each walk.

To evaluate our models' efficacy, we will use three primary metrics: MAE (Mean Absolute Error), RMSE (Root-Mean-Square Error), and Interval Score. MAE is the simplest metric and represents the sum of absolute errors divided by the sample size. This metric is intuitive, and better models will have lower values.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

RMSE is calculated similarly as a function of the difference between the expected and observed values but is more responsive to outliers due to the absolute errors being squared prior to averaging. Large differences thus incur more error than in MAE. In our experiment where we want to produce the best forecaster, disincentivizing large differences makes this popular metric crucial to consider.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Interval score (Gneiting and Raftery 2007) is used in temporal data to quantify the number of data points that fall outside of the probabilistic interval expected. Like the previous two metrics, better-performing models will have lower interval scores. IS represents the negative log -likelihood of observing a fixed-size interval around the observed values given a probability distribution.

$$\mathrm{IS}_\alpha(F, y) = (u - l) + \frac{2}{\alpha} \times (l - y) \times \mathbf{1}(y < l) + \frac{2}{\alpha} \times (y - u) \times \mathbf{1}(y > u)$$

We will use two methods to explain our model outputs: ablation and SHAP values.

One method we will use to explain our model is to do ablations on each model. To get an important feature list, we will use elastic net regularization on our dataset. We will order the feature list by importance. Using this list, we will remove features one by one and train the model. We will evaluate model performance through the metrics described early. We will plot model performance against the number of features that the model has and see what features are providing the most information.

To explain the model embeddings, our team will compute SHAP (SHapley Additive exPlanation) [8] values to get individual feature importance. The SHAP value explains how each feature changes the model prediction. The SHAP values would be able to explain model embeddings with no information on input features. We will compute the SHAP values on each model's output except for the CAMul model. For the tree-based model, Xgboost, the SHAP values can be computed using local model summarizations [9]. The CAMul model will have embedding explanations outputted for each portion of the model that indicates which features are guiding the decisions. Since each embedding has some unique properties of its own, we think there will be an interesting lesson learned through how each section acts for different features.

2.4 Expected Outcomes

The expected result of our project is a tool that can predict COVID-19 mortality after being calibrated with weekly search and mobility data from 2020 and 2021. We will then unpack the most effective model to assess the relative importance of each of the input variables. Ultimately, our project will produce important feature sets for each of the four tested models, and any overlap among them will support the heightened relevance of that specific input in generating accurate forecasters.

2.5 Remaining Work Outline

1. Train and evaluate each model (CAMul, LSTM, XGBoost, SVM) with the cleaned and joined dataset.
2. Determine the most informative aspects of each model with ablations and SHAP
3. Evaluate the lead time of each model with Pearson correlations
4. Compare the models by lead-time, effectiveness, and most informative feature sets.
5. Stretch adding CRPS (Continuous Ranked Probability Score) and Confidence Score as future metrics if time permits.

2.6 Results and Discussion Outline

- Analyze the important feature sets for each model as outputted by ablations and SHAP
- Compare and contrast the forecasting ability of the four models based on our three metrics.
- Discuss the different lead times with different models using geographic and temporal subsets.

2.7 Projected Timeline and Work Delegation

The expected deadline for this project is December 2nd, 2022. Our milestone report is due November 4th. To accomplish our midterm goals, our main focus will be the data curation, cleaning, and training one model. The data curation and cleaning will be done by Rachel in the second and third weeks of October. The last week of October will be dedicated to training one model and evaluating its results which will be done by Andy and Saideep. That will have our results for the first milestone report.

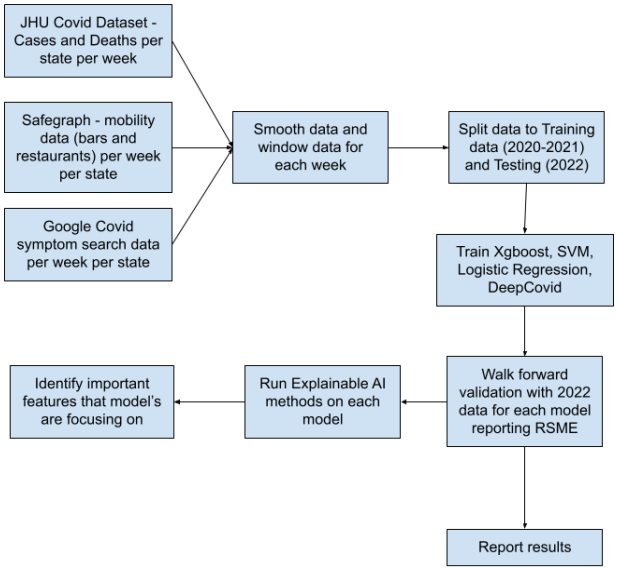


Fig 2. Flow chart of projected milestone steps.

For the second and third weeks of November, the rest of the models will be trained which will be done by Andy and Saideep. The last week of November will be dedicated to the SHAP results and visualizations which will be completed by Saideep and Rachel. All of the writing for the milestone and final report will be shared equally among the members. This project will be completed in two weeks and will cost nothing. All computing resources will come from the Emory Biomedical Informatics department.

REFERENCES

[1] James Bergstra, Daniel Yamins, and David Cox. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In Proceedings of the 30th International Conference on Machine Learning, PMLR, 115–123. Retrieved July 17, 2022 from https://proceedings.mlr.press/v28/bergstra13.html

[2] Gianfranco Cervellin, Ivan Comelli, and Giuseppe Lippi. 2017. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. Journal of Epidemiology and Global Health 7, 3 (September 2017), 185–189. DOI:https://doi.org/10.1016/j.jegh.2017.06.001

[3] Halit Cinarka, Mehmet Atilla Uysal, Atilla Cifter, Elif Yelda Niksarlioglu, and Aslı Çarkoğlu. 2021. The relationship between Google search interest for pulmonary symptoms and COVID-19 cases using dynamic conditional correlation analysis. Sci Rep 11, 1 (July 2021), 14387. DOI:https://doi.org/10.1038/s41598-021-93836-y

[4] Andrea Freyer Dugas, Mehdi Jalalpour, Yulia Gel, Scott Levin, Fred Torcaso, Takeru Igusa, and Richard E. Rothman. 2013. Influenza Forecasting with Google Flu Trends. PLOS ONE 8, 2 (February 2013), e56176. DOI:https://doi.org/10.1371/journal.pone.0056176

[5] Gunther Eysenbach. 2009. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. Journal of Medical Internet Research 11, 1 (March 2009), e1157. DOI:https://doi.org/10.2196/jmir.1157

[6] Harshavardhan Kamarthi, Lingkai Kong, Alexander Rodríguez, Chao Zhang, and B. Aditya Prakash. 2022. CAMul: Calibrated and Accurate Multi-view Time-Series Forecasting. DOI:https://doi.org/10.48550/arXiv.2109.07438

[7] Nishant Kishore. 2021. Mobility data as a proxy for epidemic measures. Nat Comput Sci 1, 9 (September 2021), 567–568. DOI:https://doi.org/10.1038/s43588-021-00127-7

[8] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. Retrieved November 4, 2022 from http://arxiv.org/abs/1705.07874

[9] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2, 1 (January 2020), 56–67. DOI:https://doi.org/10.1038/s42256-019-0138-9

[10] Stephen J Mooney, Daniel J Westreich, and Abdulrahman M El-Sayed. 2015. Epidemiology in the Era of Big Data. Epidemiology 26, 3 (May 2015), 390–394. DOI:https://doi.org/10.1097/EDE.0000000000000274

[11] Ari Seifter, Alison Rebman, Kate Geis, and John Aucott. 2010. The utility of “Google Trends” for epidemiological research: Lyme disease as an example. Geospatial health 4, (May 2010), 135–7. DOI:https://doi.org/10.4081/gh.2010.195

[12] Hu-Li Zheng, Shu-Yi An, Bao-Jun Qiao, Peng Guan, De-Sheng Huang, and Wei Wu. 2022. A data-driven interpretable ensemble framework based on tree models for forecasting the occurrence of COVID-19 in the USA. Environ Sci Pollut Res (September 2022). DOI:https://doi.org/10.1007/s11356-022-23132-3

[13] Xun Zheng, Manzil Zaheer, Amr Ahmed, Yuan Wang, Eric P. Xing, and Alexander J. Smola. 2017. State Space LSTM Models with Particle MCMC Inference. Retrieved November 4, 2022 from http://arxiv.org/abs/1711.11179

[14] Explore COVID-19 Symptoms Search Trends. Retrieved October 6, 2022 from https://pair-code.github.io/covid19_symptom_dataset

[15] Epidata API Home. Delphi Epidata API. Retrieved October 6, 2022 from https://cmu-delphi.github.io/delphi-epidata/

Identification of Optimal Features for Effective COVID-19 Disease Forecasting with Machine Learning Models

Andy Chea [1]
andychea2000@gatech.edu

Saideep Narendrula [1]
snarendrula3@gatech.edu

Rachel Calder[1]
rachelcalder@gatech.edu

1. Department of Biosciences at the Georgia Institute of Technology in Atlanta, GA, USA

ABSTRACT

The start of the COVID-19 pandemic in the age of advancing data technology allowed for the curation of data at a quality and detail level that had not previously been observed. Much of this became open-source for the collective effort of reducing the harm of the pandemic. For example, Google released data on weekly search trends per State for COVID-19-related symptoms [14]. Johns Hopkins University maintained databases of weekly data on cases, deaths, and anonymized mobility data [13]. These readily available data sets provide researchers with the opportunity to build models to have ready for future communicable disease outbreaks. Black box machine learning models are trained for the forecasting of weekly cases and deaths of COVID-19 from 2020-2022. Utilizing SHAP, important features will be extracted to evaluate the efficacy of these black box machine learning models.

CCS CONCEPTS

• Life and medical sciences • Machine Learning • Artificial intelligence

KEYWORDS

Explainable AI, Disease Forecasting, Cellular Mobility Data Tracking, Google Search Engine Trends, Xgboost, SVM, Linear Regression, SHAP

1 Background

Coined by Gunther Eysenbach in 2009 [5], "infodemiology" describes the new age of epidemiology using large swaths of open-source databases in tandem with data-mining approaches and novel data analysis to modernize the next insights in the field. This exciting era of Big Data is characterized by three "V" qualities: specific and disparate datasets (high variety), increasing numbers of samples and data points per sample (high volume), and rapid applicability and incorporation of these datums into machine learning models (high velocity) [8]. However, this final trait can sometimes clash with the first two. As the availability of new datasets balloons rapidly, how are data scientists supposed to know which ones would be most imperative to add to their forecasters? In epidemiology, where decisive action at earlier time points can alter the entire trajectory of a pandemic, it is necessary to know which portions of the torrential incoming data stream are most informative and can be used to reduce suffering immediately.

The staggering abundance of available search data represents a veritable gold mine of human experience and curiosity, and this quantity of data compounds more and more with each day. In the wake of the COVID-19 pandemic that began in 2020, Google provided a means of understanding this new public health threat: users queried how and where the virus was spreading and how to best protect themselves and their families. While many studies have since emerged around the SARS-Cov2 virus and its epidemiology, we hope to use this period of quarantine and online pandemic interest to guide the construction of models for the next communicable disease. Even now in a post-vaccine world, COVID-19 is likely to remain an endemic threat that will need to be vigilantly contained with annual boosters and public awareness campaigns.

As symptoms precede diagnoses, it follows that an individual's recognition of their own symptoms and subsequent search could be used as a less intrusive means of predicting the arrival or resurgence of COVID-19 in a population. Alongside this public availability, search data often comes anonymized which means it can be examined without complex logistics or financial compensation. While many studies throughout the years have examined symptoms as early warning signals, our project seeks to guide future work connecting symptom search data to rapid, actionable epidemiological insights by discovering informative variables.

1.1 Literature Review

When predicting waves of infection, epidemiologists can factor in mobility data and use the assumption that the length of time of proximity would increase the likelihood of new cases. However, mobility data should not be solely considered because it does not indicate other protective measures individuals could be using. For example, data from those in close proximity with preventative measures in place such as mask usage could cause an overestimate in

the prediction of cases [6]. Mobility data can also vary based on the demographics of the individuals surveyed, as the individuals must have access and exposure to the technology used. Therefore, the data must be considered to be a sample of the total set.

The relationship between Google search trend data and COVID-19 cases has also been evaluated for predictive power. The assumption is that individuals may be more inclined to search for COVID-19 symptoms if they or someone they came into contact with are experiencing symptoms, and that search results could be used as a pseudo-diagnostic tool. A study using Google search trends in Turkey, Italy, Spain, France, and the United Kingdom found there to be significant time-varying correlations between pulmonary symptom search trends and new cases [3].

Current literature using search data to make epidemiological predictions has had mixed success. A pioneering paper from 2010 by Seifter et al [11] evaluated the effectiveness of the newly-released Google Trends database in mimicking seasonal Lyme disease trends. While rather simple twelve years later, this paper seemed to indicate then that search data was a useful resource in modeling. The authors also noted that the states where Lyme disease was considered endemic were also often the ones with the highest search quantity for terms like "Lyme disease" and "tick bites".

Then in 2013, Dugas et al [4] used the more specialized Google Flu Trends to train several forecasting models that predict the hospital load of flu patients in the following week. Their analysis was steeped in applicability as they focused only on data that was collectible in real-time and that could inform medical institutions at many levels. The most successful model was a generalized linear autoregressive moving average method (GARMA) with a negative binomial distribution. With a leave-one-out validation approach of 7 sets, they recorded an accuracy of 83%, lending credence to the validity of search data in our analysis. One caveat that they admit to is using only case data from one medical center, which might have overspecified the model. We will avoid this by examining cases from many different states to train a more generalizable model.

In 2017, a skeptical paper from Cervellin et al [2] claims that while previous work had found some correlation between search and epidemiological trends, there is much more nuance to consider before claiming these findings are substantial. Their experimental design found that both common diseases of low media coverage and rare, sensational diseases did not correlate well with their population prevalence. However, their methodology involves some very specific terms that even someone with the symptoms might not know to use when looking for more info, such as "renal colic" and "epistaxis." While all search terms are undoubtedly affected by media influence, these words might be too far removed from the average user to present a relevant test case. Our project utilizing symptom searches, which come more naturally than medical terminology, should prove to be informative.

In terms of prediction approaches, traditional machine learning methods have been used to compare against epidemiological models. In 2022, Rahman et al. [9] compared the accuracy of Xgboost compared to the ARIMA model with COVID-19 data in Bangladesh. In their approach, they found that ARIMA performed better than Xgboost in predicting cases and the number of deaths. With this approach, there was no explainability with their model so it was hard to see what features their Xgboost model was focusing on. There have been approaches to predicting cases and deaths of Covid 19 with ensemble models and explaining their approach with SHAP via Zheng et al [12] in 2022. Their approach utilized input features such as self-protection, policy indicators, community movement, and the time index. The SHAP toolbox they used provided information on the importance of the input features.

2 Proposal

2.1 Problem Statement

Simply put, our goal is to discover the symptom search terms that are most informative in training effective disease transmission machine learning models. We will integrate state-level COVID-19 symptom search datasets with mobility and case datasets from each week in 2020 and 2021 to train a model that predicts COVID-19 mortality in 2022. This walk-forward approach will involve retraining the model while considering the previous month's incoming data. Extracting the most informative variables from the most updated model can then help inform other research groups about what types of data from our aggregated set are most vital to include in their forecasting tools.

2.2 Datasets

The Johns Hopkins University database of confirmed COVID-19 cases will be used to capture weekly cases and deaths within each American state from 2020 to 2021 and will be accessed through the Delphi Epidata API [13]. The Delphi Epidata API will also be used to query data from the SafeGraph database based on weekly metrics per state.

The SafeGraph database contains anonymized cell phone location data until April 19th, 2021 with flags for highly populated locations such as bars and restaurants.

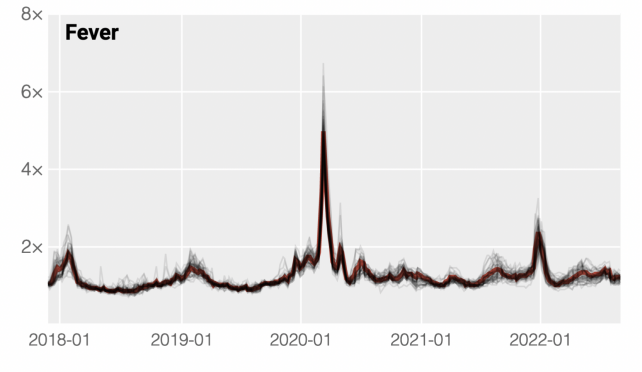


Fig 1. Aggregated Fever-related search term frequency in the United States per state over time. The red line represents an average of state search queries [14].

The Google COVID-19 Symptoms Search Trend database will be used to identify counts of google searches mapped back to COVID-19 symptoms [14].

2.3 Methodology

Our approach will utilize machine learning models to evaluate the cases and deaths from Covid-19 in each state per week. The mobility and symptom data will be condensed into weekly data to match the JHU weekly reporting of the deaths and cases. For each week-length window, the data will be smoothed using Locally Weighted Scatterplot Smoothing (LOWESS) [10] and statistics will be computed on each window. Since the JHU case data is already reported at the weekly level, we will utilize LOWESS to smooth the data.

Xgboost, SVM, LGBM, and Linear Regression will be used as the machine learning models. Each model can be implemented with the LGBM, Xgboost, and sci-kit learn packages. To optimize the model’s hyperparameters, Bayesian Optimization will be performed to identify the best set of parameters utilizing hyper-opt [1]. These models can be used with the SHAP [7] toolbox which will provide information on the contribution of each input feature.

To train each model, we will use a train-validation approach. The 2020-2021 data will be used to train the model. To validate each model, we will use a walk-forward validation approach every two months in the 2022 data and average the results from each walk. The metrics we will be measuring to evaluate the efficacy of the model will be Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), Mean Percentage Error (MPE).

2.4 Expected Outcomes

The expected result of our project is a tool that can predict COVID-19 mortality after being calibrated with weekly search and mobility data from 2020 and 2021. We will then unpack the most effective model to assess the relative importance of each of the input variables. Ultimately, our project will produce both a machine learning model and an explanation for why it is effective, which will be useful for future groups and those working with other diseases.

2.5 Evaluation

To evaluate our progress, our mid-term check will be the exploratory data analysis on the input and output data. We would also see how well the linear regression model is performing on this dataset. For our final report, we would expect to see the metrics on each model compared to each other. Our SHAP plots would show us the important features as well for each model.

2.6 Projected Timeline and Work Delegation

The expected deadline for this project is December 2nd, 2022. Our milestone report is due November 4th. To accomplish our midterm goals, our main focus will be the data curation, cleaning, and training one model. The data curation and cleaning will be done by Rachel in the second and third weeks of October. The last week of October will be dedicated to training one model and evaluating its results which will be done by Andy and Saideep. That will have our results for the first milestone report.

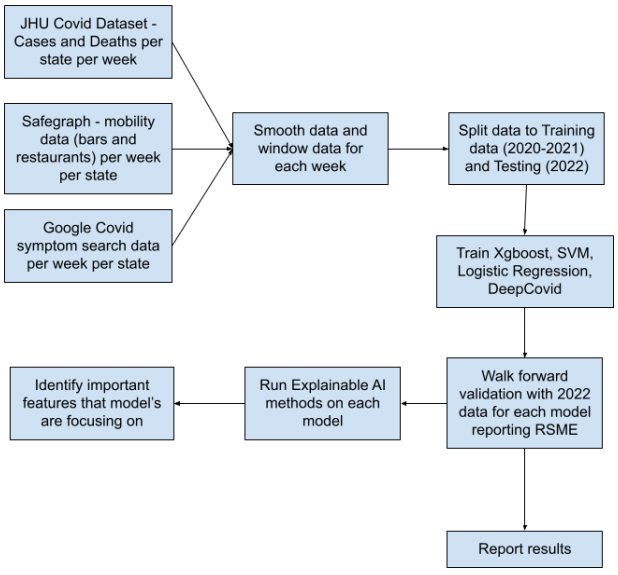


Fig 2. Flow chart of projected milestone steps.

For the second and third weeks of November, the rest of the models will be trained which will be done by Andy and Saideep. The last week of November will be dedicated to the SHAP results and visualizations which will be completed by Saideep and Rachel. All of the writing for the milestone and final report will be shared equally among the members. This project will be completed in two weeks and will cost nothing. All computing resources will come from the Emory Biomedical Informatics department.

REFERENCES

[1] James Bergstra, Daniel Yamins, and David Cox. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on Machine Learning*, PMLR, 115–123. Retrieved July 17, 2022 from https://proceedings.mlr.press/v28/bergstra13.html

[2] Gianfranco Cervellin, Ivan Comelli, and Giuseppe Lippi. 2017. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. *Journal of Epidemiology and Global Health* 7, 3 (September 2017), 185–189. DOI:https://doi.org/10.1016/j.jegh.2017.06.001

[3] Halit Cinarka, Mehmet Atilla Uysal, Atilla Cifter, Elif Yelda Niksarlioglu, and Aslı Çarkoğlu. 2021. The relationship between Google search interest for pulmonary symptoms and COVID-19 cases using dynamic conditional correlation analysis. *Sci Rep* 11, 1 (July 2021), 14387. DOI:https://doi.org/10.1038/s41598-021-93836-y

[4] Andrea Freyer Dugas, Mehdi Jalalpour, Yulia Gel, Scott Levin, Fred Torcaso, Takeru Igusa, and Richard E. Rothman. 2013. Influenza Forecasting with Google Flu Trends. *PLOS ONE* 8, 2 (February 2013), e56176. DOI:https://doi.org/10.1371/journal.pone.0056176

[5] Gunther Eysenbach. 2009. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. *Journal of Medical Internet Research* 11, 1 (March 2009), e1157. DOI:https://doi.org/10.2196/jmir.1157

[6] Nishant Kishore. 2021. Mobility data as a proxy for epidemic measures. *Nat Comput Sci* 1, 9 (September 2021), 567–568. DOI:https://doi.org/10.1038/s43588-021-00127-7

[7] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2, 1 (January 2020), 56–67. DOI:https://doi.org/10.1038/s42256-019-0138-9

[8] Stephen J Mooney, Daniel J Westreich, and Abdulrahman M El-Sayed. 2015. Epidemiology in the Era of Big Data. *Epidemiology* 26, 3 (May 2015), 390–394. DOI:https://doi.org/10.1097/EDE.0000000000000274

[9] Md Siddikur Rahman, Arman Hossain Chowdhury, and Miftahuzzannat Amrin. 2022. Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh. *PLOS Global Public Health* 2, 5 (May 2022), e0000495. DOI:https://doi.org/10.1371/journal.pgph.0000495

[10] William R. Schucany. 1995. Adaptive Bandwidth Choice for Kernel Regression. *Journal of the American Statistical Association* 90, 430 (June 1995), 535–540. DOI:https://doi.org/10.1080/01621459.1995.10476545

[11] Ari Seifter, Alison Rebman, Kate Geis, and John Aucott. 2010. The utility of “Google Trends” for epidemiological research: Lyme disease as an example. *Geospatial health* 4, (May 2010), 135–7. DOI:https://doi.org/10.4081/gh.2010.195

[12] Hu-Li Zheng, Shu-Yi An, Bao-Jun Qiao, Peng Guan, De-Sheng Huang, and Wei Wu. 2022. A data-driven interpretable ensemble framework based on tree models for forecasting the occurrence of COVID-19 in the USA. *Environ Sci Pollut Res* (September 2022). DOI:https://doi.org/10.1007/s11356-022-23132-3

[13] Epidata API Home. *Delphi Epidata API*. Retrieved October 6, 2022 from https://cmu-delphi.github.io/delphi-epidata/

[14] Explore COVID-19 Symptoms Search Trends. Retrieved October 6, 2022 from https://pair-code.github.io/covid19_symptom_dataset