Estimating Vaccination Uptake for COVID-19 with a Focus on Interpretability

Kartik Narang
Georgia Institute of Technology
Atlanta, Georgia
knarang31@gatech.edu

Amish Saini
Georgia Institute of Technology
Atlanta, Georgia
asaini68@gatech.edu

ABSTRACT

The COVID-19 pandemic has had a devastating impact worldwide, causing regional disparities in COVID-19 prevalence to emerge. These disparities demonstrate a need for predictive modeling strategies to make the best decisions that reduce the burden on our healthcare system. There were many periods where the vaccination rates in the US were stagnant, despite widespread availability at many points. Understanding the most important factors that impact vaccination uptake using Genetic Programming (GP) and interpretable Tree-Based models can be fruitful to increase vaccination coverage.

Formally, we apply GP and Tree-Based models to forecast COVID-19 vaccination uptake and identify the most important factors that cause stagnant periods of vaccination rates. The goal with this is that public health officials can develop targeted interventions for groupings of populations to promote vaccine uptake.

KEYWORDS

Genetic Programming, Evolutionary Algorithm, Epidemiology, Vaccination, COVID-19, Forecasting, Random Forest, XGBoost, AdaBoost

ACM Reference Format:
Kartik Narang and Amish Saini. 2023. Estimating Vaccination Uptake for COVID-19 with a Focus on Interpretability . In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Amidst the challenges posed by the COVID-19 pandemic, numerous research facilities successfully created safe and efficient vaccines at an unprecedented pace. Nevertheless, within affluent nations, a segment of the population ranging from 5% to 30% has yet to receive a single COVID-19 vaccine dose, and even larger portions of people in lower-income countries remain unvaccinated [7]. The rapid and extensive vaccination against COVID-19 contributes significantly to mitigating the severity of the disease, as well as curbing the transmission of the virus. Consequently, resistance to, deferral of, or the inability to participate in vaccination programs presents a substantial societal challenge. Identifying individual factors that influence vaccine uptake can empower policymakers to develop more targeted and effective interventions for future immunization initiatives.

Policymakers have been trying to find the root cause of the stagnant vaccination rates. Although granular region-based analysis is sparse, there are discussions over the aggregate population that hint towards potential causes. For example, the COVID-19 vaccine was developed and rolled out very quickly, which led to concerns about safety and efficacy among some people. Additionally, the COVID-19 pandemic has been highly politicized, which has further contributed to vaccine hesitancy. There is also concern about trust within communities and the spread of misinformation. Misinformation and disinformation about vaccines can spread rapidly online and through social media. This misinformation can lead to fear and distrust of vaccines and can discourage people from getting vaccinated. Furthermore, people who have had negative experiences with the healthcare system in the past may be less likely to trust healthcare providers and public health officials. This is especially true for people of color and other marginalized groups.

There is also the issue of complacency. Complacency can set in when regions experience declining infection rates, leading some individuals to perceive vaccination as unnecessary. The success of vaccination campaigns can inadvertently breed a false sense of security, making it difficult to motivate people to get vaccinated until another surge in cases occurs. This cyclic pattern hampers long-term vaccination efforts.

2 RESPONSE TO MILESTONE COMMENT

We iterated on the following feedback from the midpoint milestone. The first suggestion was that we should add more statistical analysis of the data in addition to the correlation analysis, and say the histogram of each feature mean/variance analysis. We did this but kept it for our appendix in Figure 12. We also conducted cross correlation analysis in Figure 3. The second feedback point was to fix the caption for Figure 1, which had a placeholder caption. This has now been fixed to reflect what the chart actually represents. We were also told to have 15 or more references in the final report, and we increased our background research and incorporated 16 references in this iteration. Another suggestion was to illustrate more on XGBoost or AdaBoost, and we did this by describing our experiments thoroughly and plotting multiple forecasts and statistics. The last feedback is that we should elaborate and focus on interpretability, and we did this by using SHAP as explained in section 6.1 and showed in Figures 9 and 10.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 RELATED WORK

Previous work has been explored to predict vaccination uptake. Some of the first were in the case of Influenza, where Shaham et al. compared Logistic Regression, Naive Bayes, XGBoost, and Artificial Neural Networks (ANNs) to predict whether a patient would get vaccinated in the near future [13]. Cheong et al. gained inspiration from this and focused on XGBoost and Sociodemographic data to apply similar methods to COVID-19. However, the results are quite poor, predicting vaccination uptake across US counties with 62% accuracy [2].

In an effort to improve these results, Sigalo et al. [15] investigated augmenting data from Social Media, in this case Twitter, to understand the attitudes found in tweets in hope of using it as an additional feature on top of six months of publicly available COVID-19 data from the CDC. They found that this augmentation reduced root mean square error by as much as 83%. However, one big limitation of this study was that a lot of populations weren't considered, as limiting the study to the collection of Tweets only in English could lead to a biased understanding of sentiments, since populations that don't speak English as their native language are excluded. From this work, Sattar and Arifuzzaman conducted a study where they used a time-series model to project that by the end of July 2021, 62.44% of the US population would get at least 1 dose of the COVID-19 vaccine [12]. However, this study suffers from the fact that they used aggregate data over the whole nation instead of more granular geographic data. Additionally, this study falls to the limitation of common time-series models - a lack of interpretability.

There has also been some initial work with search trends to predict disease epidemiology in the United States. Mattiuzzi and Lippi [8] investigated search trends related to 16 symptoms and they found that the combination of these symptoms explained 88% of total variance in COVID-19 Google searches. However, there's a big factor that this paper doesn't account for, which is that there's often a lag time between the search and when action is actually taken, such as reporting the vaccine or getting vaccinated. This is something we consider in our proposed solution. Rovetta [10] used Google Search Trends to predict vaccinations in Italy. The methodology used is substantial and we gain inspiration from it, but one critical downfall is that the results aren't interpretable because ARIMA, which is what the paper uses, doesn't give a way to determine which feature caused the observed trend. This paper introduces a lag associated with the Google Search Trends. Hassani and Silva [5] further that this is a critical issue when implementing a novel forecasting approach for energy using Google trends.

To motivate our solution, we look at two papers to drive some of our mobility features. Shortall et al. [14] do a wholisitic review on literature on the COVID-19 measures for passenger mobility and propose a research agenda while discussing the policy relevance of their findings. From this paper, we deduce that it would likely be useful to include metro, bicycle, bus, and tram transit data if accessible since the authors find that they were a significant factor in the spread of the disease. Müller et al. [9] further confirms the importance of these features by creating human mobility trace models and confirming that these hubs were significant sources of spread of the disease.

Despite their use in vaccination forecasting, Genetic Programming and symbolic regression have shown previous success in epidemiological forecasting, specifically for COVID-19. Anđelić et al. [1] used GP to evolve mathematical equations estimating confirmed, deceased, and recovered COVID-19 cases in several countries. By combining these country-level equations, they generated models forecasting the epidemiology curves and active cases over time. The GP-derived models provided a close fit to the real epidemiological data.

In another study, Salgotra et al. [11] developed GP models to predict daily confirmed and death cases due to COVID-19 in Indian states. The explicit mathematical equations produced by GP achieved high accuracy in estimating the real data and enabling reliable short-term forecasts. Both studies demonstrate GP's capability for symbolic regression to produce simplified yet accurate epidemiological forecasting models without extensive data requirements. The white-box nature of the models provides insights into disease spread. Such techniques could be leveraged to model other epidemiological phenomena such as vaccination uptake.

Most applied intelligence methods that have constructed time-series models for vaccinations are impossible to interpret regarding their factor's impacts. Beyond this, GP and Tree-Based techniques intrinsically have an implicit process of feature selection since they can employ more interpretable representations [3]. Therefore, we use Genetic Programming (GP) and Tree-Based models to develop models and expressions that forecast the COVID-19 vaccination uptake that can be taken apart to understand the factors that impact vaccination uptake. We do so by comparing the best expressions that we get from GP to assess the most influential and common features, and for Tree-Based models, SHAP tree-explainers and permutation importance are the primary methodology.

4 DATA

We use a joined dataset from the Google COVID-19 Open Data Repository [16] sourced by numerous agencies which contains daily information about variety of topics regarding COVID-19. Specifically, the joined dataset is an accumulation of various datasets within the repository including Vaccinations, Vaccinations Search, Search Trends, Epidemiology, and Mobility datasets. The joined dataset contains data from January 1st, 2020 to September 15th, 2022 in .csv format, and we focused on the state of Georgia. The data is publicly available to download for free. The joined dataset is indexed based on the date and region. Our columns of interest include daily vaccination counts, vaccination search counts, symptom search counts, case counts, test counts, and mobility metrics. An additional source that we looked at was the Google Trends website, which gave us weekly statistics for searches related to vaccinations or where to find them and COVID in Georgia. The vaccination counts are our target while the remaining columns are available for Symbolic Regression and Tree-Based Regressors to forecast.

5 METHODOLOGY

Our code that implements the following methodologies can be found in the Appendix.

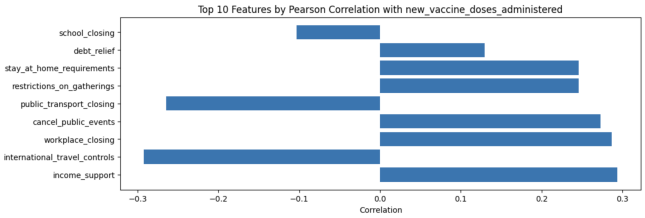

Figure 1: Displaying Top 10 Features and their Corresponding Correlation via Pearson Correlation

5.1 Exploratory Data Analysis

For EDA, we use Pearson Correlation, Spearman Rank, and Mutual Information to determine how important the features are and if we should remove or alter them. Pearson Correlation Coefficient gives us a value between -1 and 1 that tells the strength and direction of the relationship between two variables, in this case with each feature against the target variable. We use this to figure out which features we should drop. Similarly, Spearman's rank correlation coefficient also ranges from -1 to 1, but it can evaluate monotonic relationships whereas Pearson Correlation only assesses linear relationships, so we run both to get more comprehensive results. Lastly, we use mutual information to give us similar insights, where a value further away from zero indicates more association between the two variables and a value closer to zero indicates less association between the two variables.

We grouped all of our features into 7 categories: cdh, cdh_age, pop, mob, misc, search, and weather. 'cdh' stands for confirmed, deceased, and hospitalized, so we grouped any of the features in that list. 'cdh_age' does this by age. Population, mobility, miscellaneous, search trends, and weather are among the other columns. For each of these groupings, we calculate summary statistics (mean, std, min, max, etc.) for the columns in the current group and we plot histograms to visualize the distribution for each column's data. We then calculate a correlation matrix as a heatmap which visually represents the correlation values between columns in the current group.

A couple of things that we noticed was that the population was unchanging and weather wasn't the most important factor, while cdh-related columns deemed higher importance. Then, we run Pearson Correlation, Spearman Correlation, and Mutual Information between features in the current group and the label feature and display horizontal bar charts for the top 10 features based on each correlation measure. In Figure 1, we see that income support and workplace_closing have high correlation with vaccination uptake, whereas there is a reverse correlation for public transport and international travel control since shutting down public transport and international travels likely made people feel less of a need to get vaccinated since they wouldn't be traveling. We do this to figure out which features actually matter (feature selection) because those that perform well are indicative features in our model.

Another factor that we noticed was that our features, such as mobility, hospitalizations, Google Search Trends, etc. doesn't impact vaccination uptake immediately. For example, the time that someone searches a symptom or a keyword related to the vaccine

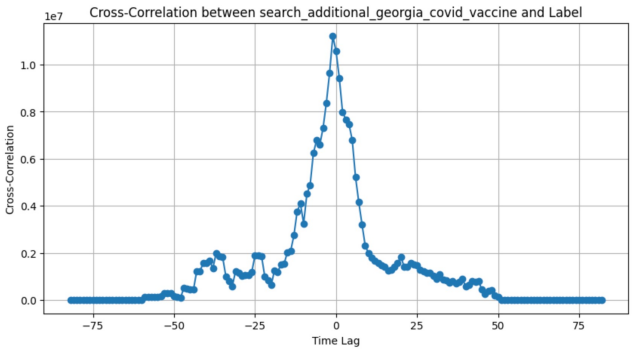

Figure 2: Cross-correlation Analysis for the Google Search for (Georgia Covid Vaccine)

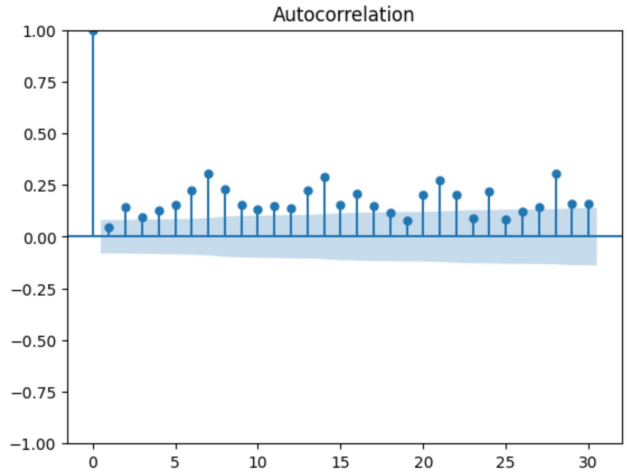

Figure 3: Autocorrelation to Show Weekly Spikes

doesn't necessarily correlate to when they got the vaccine. Therefore, we would have to account for this lag using Cross-correlation Analysis. Cross-correlation analysis inherently identifies the time lag between two time series signals, helping understand the temporal relationships between them. By analyzing the cross-correlation function, we can pinpoint the time lag at which specific features most strongly correlate with changes in vaccination uptake. For example in Figure 2, we see that the time lag for when people search about the Georgia COVID Vaccine is -1. This means that this search has a 1 day lead time in comparison to the vaccination uptake.

Finally, we looked at our label, the number of daily vaccines. We noticed that the values seem to be updated at the end of each week. In order to confirm this, we used an autocorrelation plot as shown in Figure 3 to depict spikes at 7, 14, 21, and 28. This indicates that our label is likely updated weekly. Therefore, we change our label to number of weekly vaccinations moving forward.

After data collection and feature engineering, we implemented Genetic Programming through symbolic regression and Tree-Based regressors such as Random Forest, XGBoost, and ADABoost to forecast the number of vaccinations administered per day. For our

two subdomains of models, different approaches are used for interpretability.

5.2 Tree-Based Regressors

The models that we used to fit our data were Tree-Based models, namely XGBoost, AdaBoost, and Random Forests.

XGBoost, or Extreme Gradient Boosting, works by training a number of decision trees, with each tree trained on a subset of the data, and the predictions from each tree are combined to form the final prediction. It combines a number of weak learners to form a strong learner. Although each weak learner is only slightly better than random guessing, when combined together, they form a strong learner which is much more accurate. In this case, the weak learners are regression trees, and each of these trees maps an input data point to one of its leaves that contains a continuous score. It's called gradient boosting because of its use of gradient descent to minimize the loss when adding new models to the ensemble.

AdaBoost, or Adaptive Boosting, also works on the principle of adding together multiple weak learners to get strong learners. It punishes incorrectly predicted samples by assigning a larger weight to them after each prediction round. Because of its relative simplicity and interpretability, we use it as a baseline to compare against XGBoost and Random Forests. It tends to be less prone to overfitting as the input parameters are not jointly optimized.

Random Forests is based on the bagging approach, where it grows out multiple decision trees that are merged together for a more accurate prediction. While each tree in a Random Forest is constructed using a random subset of training data and a random subset of features at each split, methods like XGBoost and AdaBoost build out the trees sequentially. This randomness helps in reducing overfitting and improving generalization.

5.3 Genetic Programming

We used symbolic regression, which is a subset of Genetic Programming. The way symbolic regression works is that it searches over a space of all possible mathematical formulas and their transformations, starting from a basic set of functions and finding the ones that best predict the output variable.

The baseline algorithm is an Evolutionary Algorithm (EA) as shown in Figure 3. We started with a population composed of randomized expressions and seeded well-known models which we can call our original set of individuals. These individuals are expressed in a tree format to allow for the possibility of mating and mutations. Every individual is deemed a fitness value based on their performance on the test data. Fitness for our example is multi-objective using our 3 metrics: RMSE, MAE, and R^2. The highest-fitness individuals are selected to be a part of a mating pool from which they can mate. Mating two tree individuals can be done in many ways, with each mating method having its own probability of occurring. Mutation is similar in its variety of methods with each having its own probability. However, mutation can occur to any member of the population. In order to prevent overpopulation, lower-fitness individuals are removed based on a threshold. Over many generations, our artificial system of natural selection and gene variation generates an expression that models the curve for vaccination uptake. The implementation of this process is done using the DEAP library [4].

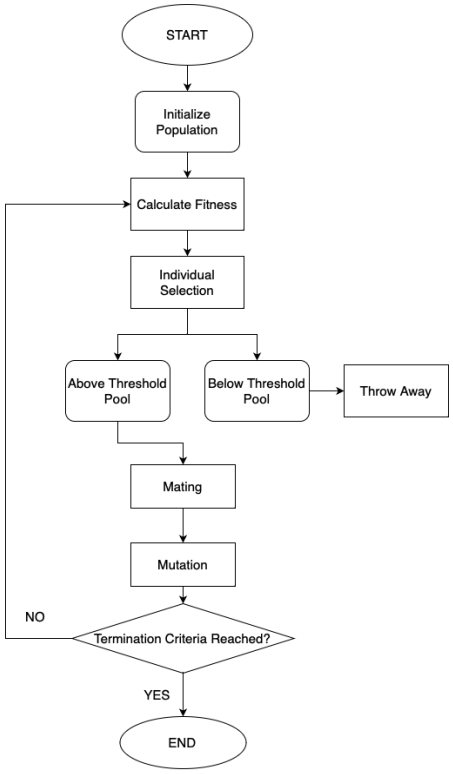

Figure 4: The Evolutionary Algorithm Framework used to generate regression expressions from Genetic Programming.

5.4 Interpreting Results

Our choice of using GP and Tree-Based regressors comes from our motivation to understand the interworkings of our successful expressions and models. Specifically, what factors have the greatest impact on vaccination uptake so that policy-makers can understand how their decisions and environment impact uptake and what actions they can make to increase vaccination uptake.

In our analysis of Tree-Based regressors, we leverage the power of SHAP (SHapley Additive exPlanations [6]) with a tree explainer to assess feature importance and feature impacts within ensemble models, specifically Random Forest, XGBoost, and Adaboost. SHAP's tree explainer technique offers a systematic way to understand how each feature contributes to the model's predictions. Utilizing Shapley values, we can quantify the average impact of each feature across all possible feature permutations. This comprehensive approach sheds light on feature interactions and their influence on model predictions. By visualizing SHAP values in summary plots and instance-level explanations, we gain valuable insights into the relative importance of each feature. This analysis enables us to make informed decisions regarding feature selection and enhance our understanding of these ensemble models, ultimately increasing their transparency and trustworthiness.

Analyzing Genetic Programming expressions to identify the most important features involves several potential techniques and strategies. One method is to examine the presence and frequency of features within evolved expressions. Features that consistently appear in symbolic equations across different runs of the Genetic Programming algorithm are indicative of their significance. Additionally, evaluating the relative contribution of each feature to the overall fitness of the symbolic expressions can reveal their importance. This can be done by measuring the impact of feature removal on the model's performance. We can control the features that are used and seed populations with different features to compare performances of the best individuals.

5.5 Model Evaluation

Our expressions are generated using symbolic regression and models are created using Tree-Based regressors trained on the vaccination uptake data and the given features. The expressions and model outputs are evaluated using Normalized Root Mean Square Error (NRMSE), Normalized Mean Absolute Error (NMAE) and R-squared. These metrics are calculated based on the comparison of the test set of the actual data and our expressions' predictions. Evaluation for the Genetic Programming expressions occur during the genetic simulation and are encoded into the fitness function. The highest-fitness individual has the lowest NRMSE, NMAE, and highest R-squared. For our Tree-Based regressors, these metrics are applied to their predictions once training is finished.

6 EXPERIMENTS/RESULTS

The following questions are to be answered using our Tree-Based Regressors and Genetic Programming as mentioned previously.

We aim to answer the following questions:

- Are we able to accurately predict vaccination uptake?
- Can we understand the features that impact these predictions in order to provide suggestions to health officials?

We intend to answer these questions differently for Tree-Based Regressors versus Genetic Programming. Regarding Tree-Based Regressors, we train on a subset of data and attempt to forecast the following weeks. Using SHAP, we analyze our forecast to understand how much impact each feature has on the model's prediction. This allows us to understand what exactly leads to an increase or decrease in vaccination uptake. Regarding Genetic Programming, we use our Genetic Algorithm from section 5.3 to come up with a symbolic regression expression that attempts to fit the entire vaccination uptake curve. After doing so, we can observe the most influential factors based on their appearance in successful symbolic expressions.

6.1 Tree-Based Regressors

At first, we attempted to use a rolling window technique for our trained and test sets for each of our Tree-Based Regressors. This is how the process works: The time series data is split into consecutive, non-overlapping folds or segments, where each fold represents a contiguous block of time. The model is then trained on the data from the beginning up to a certain point in time. The model is tested on the subsequent time period. This process is repeated multiple times, with the training and testing sets advancing in time for each iteration.

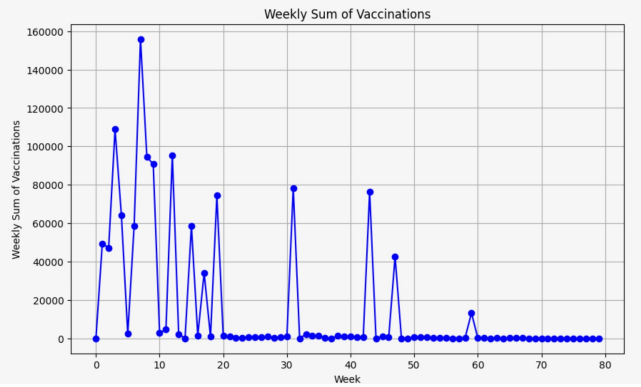

Figure 5: Weekly Sum of Vaccinations: Initial Dates have Larger Density of Uptake Spikes

Table 1: Original Tree-Based Regressor Metrics

Model	Norm. MAE	Norm. RMSE	R-Squared
Random Forest	15.44	16.45	-15850.82
XGBoost	5.44	7.69	-3027.03
AdaBoost	3.40	3.59	-610.25

However, we got terrible results using this approach as shown in Table 1. Specifically, the Normalized RMSE and Normalized MAE were high and the R^2 values were extremely negative. Therefore, we took a look back into our data and realized that the majority of the vaccination uptake spikes would be contained in the training portions as shown in Figure 5. Therefore, our models are biased to expecting uptake spikes in the forecast. In order to tackle this, we changed the focus of our model to being able to forecast future uptake spikes.

To forecast these uptake spikes, we used various train and test splits. For example, we trained on the spikes from weeks 0-14 in order to forecast the spike on week 19. These training ranges were also expanded to then predict the spikes on weeks 31 and 43.

This experiment was far more successful as shown in Figure 6. The top plot in the figure correctly forecasts that the vaccination uptake rapidly increases after week 18 into the spike that is shown in the red oval in the bottom plot. Similar success is shown in Figure 7 for week 31 and in Figure 8 for week 43. The metrics for these 3 spike forecasts can be found in Table 2. We can see that the normalized RMSE and normalized MAE values are quite low, and the R^2 is decently high. This validates that our forecasts are much better than before.

After our successful forecasts, we generated SHAP values in order to understand the variable impact that led to these spikes. After generating SHAP values, we plotted the feature impacts for specific spike forecasts. The length of each red subsection corresponds to the magnitude of impact that feature had on the model's final prediction to be high. Conversely, blue corresponds to the impact

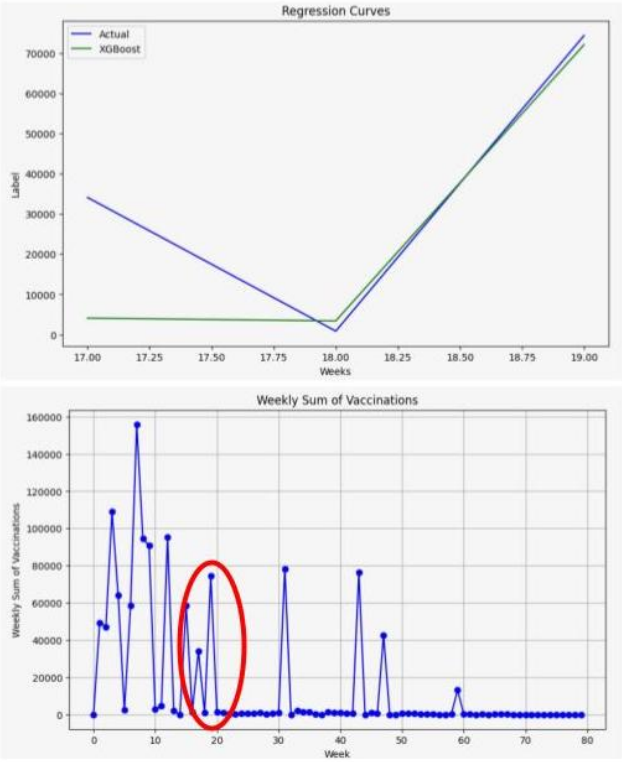

Figure 6: XGBoost Regressor Forecast Predicting Spike Starting Week 18

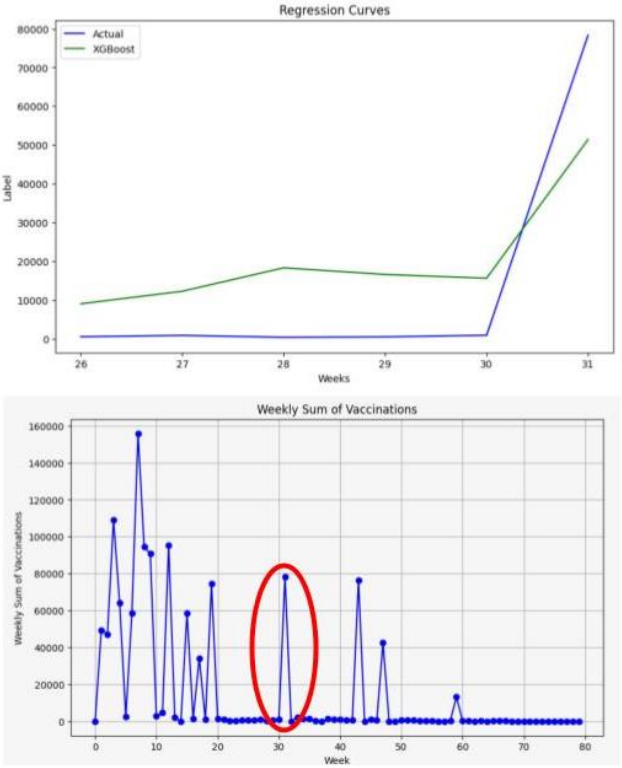

Figure 7: XGBoost Regressor Forecast Predicting Spike Starting Week 30

Table 2: Spike Forecast Tree-Based Regressor Metrics

Model Type	Spike Start Week	Norm. RMSE	Norm. MAE	R-Squared
XGBoost	18	0.2373	0.1580	0.6631
XGBoost	30	0.2175	0.2044	0.6572
AdaBoost	42	0.2291	0.1153	0.6204

for a lower prediction. In Figure 9, we can see that the feature that has the greatest impact on the model's decision to forecast a spike is the number of new hospitalized patients between the ages of 50 and 60. Specifically, the number of hospitalized patients within that age range was 36 for that week. Interpreting this impact is a little more tricky because despite the fact that the number of hospitalized patients may impact vaccination uptake, the spikes in our data are sparse. However, one may be able to interpret this in many ways. For example, people may perceive a higher risk of contracting the disease if there is an increase in the number of hospitalized patients. This perceived risk can motivate individuals to seek vaccination as a preventive measure to reduce their chances of getting seriously ill.

We can analyze a different spike in Figure 10. In this plot, we see that the most impactful feature in the model's prediction of a spike is the search trend for hyperventilation. This may be due to symptom awareness, since individuals experiencing symptoms associated with hyperventilation may be prompted to search for related information. As a result, individuals may become more aware of COVID-19 and its potential severity, leading them to consider vaccination.

The sparse representation of features in a limited dataset can lead to less accurate SHAP values, as the model may struggle to capture the nuanced relationships between features and predictions. The resulting instability and variability in model behavior, coupled with increased sensitivity to outliers, further compromises the robustness of SHAP values. Additionally, the risk of overfitting is heightened in scenarios with scarce data, potentially causing SHAP values to be influenced by noise rather than true patterns.

6.2 Genetic Programming

In order to fit the vaccination curve and interpret our expressions, we began by running various experiments on our Genetic Algorithm explained in Section 5.3. Specifically, we altered the following factors: number of features, number of generations, population size, mating probability, individual tree depth, and mutation probability. Through our various experiments, we found that increasing the size of the population and number of generations tended to give us better results for symbolic expressions. However, we failed to reach a point in any of our experiments where our generated symbolic expression fit the data well using only the features given. Our best symbolic regression expression is shown in Figure

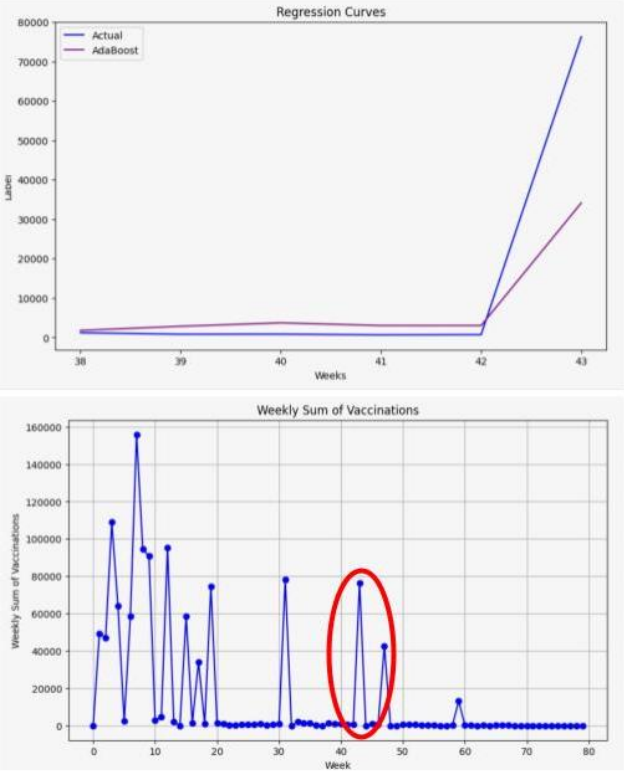

Figure 8: XGBoost Regressor Forecast Predicting Spike Starting Week 42

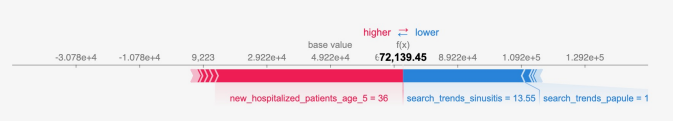

Figure 9: Interpreting Feature Importance with SHAP for Week 18

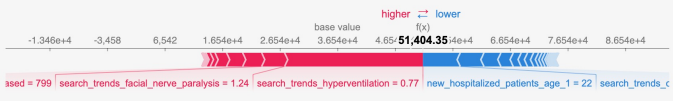

Figure 10: Interpreting Feature Importance with SHAP for Week 30

Due to our failure in fitting vaccination uptake well, analyzing our top expressions for their feature importance became insignificant. Our failures are likely a combination of the following reasons. First, predicting vaccination uptake using factors like mobility, search trends, and hospitalizations alone is challenging due to the complexity of individual decision-making. Next, Genetic Programming requires an extreme amount of computational power. As mentioned before, our results got better if we increased the number of generations and the size of our population. However, an

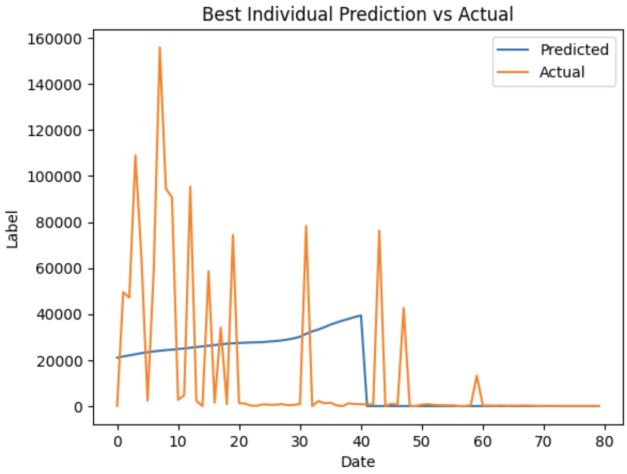

Figure 11: Best Symbolic Regression Expression against Actual

increase in both of these values caused an exponential increase in the runtime of these algorithms.

7 CONCLUDING DISCUSSION

From our literature review, we found that predicting vaccination uptake is a difficult task. Previous research had only managed to be successful using models such as ARIMA, which are not interpretable and provide less information to help officials make decisions. Using Tree-Based Regressors, we were able to forecasts peaks with sufficient accuracy and interpret the feature impacts that led to our models predictions. Genetic Programming failed due to the difficulty of fitting an entire curve and the bottleneck of computational complexity. Additionally, it is important to note that our interpretability is decreased and loses much of its validity due to the scarcity of data.

In terms of next steps, we want to use Tree-Based Regressors to forecast on a disease that has data for a much longer time period, like the Flu, in order to get more confident predictions and interpretations.

REFERENCES

[1] N. Anđelić, S.B. Šegota, I. Lorencin, and et al. 2021. Estimation of covid-19 epidemiology curve of the United States using genetic programming algorithm. *International Journal of Environmental Research and Public Health* 18, 3 (2021), 959.

[2] Q. Cheong, M. Au-yeung, S. Quon, K. Concepcion, and J.D. Kong. 2021. Predictive modeling of vaccination uptake in US counties: A machine learning–based approach. *National Library of Medicine* (2021).

[3] P.G. Espejo, S. Ventura, and F. Herrera. 2010. A survey on the application of genetic programming to classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40, 2 (2010), 121–144.

[4] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. 2012. DEAP: Evolutionary Algorithms Made Easy. *Journal of Machine Learning Research* 13 (jul 2012), 2171–2175.

[5] H. Hassani and E. Silva. 2016. Forecasting energy data with a time lag into the future and google trends. *International Journal of Energy and Statistics* 04, 04 (2016), 1650020. https://doi.org/10.1142/s2335680416500204

[6] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR* abs/1705.07874 (2017). arXiv:1705.07874 http://arxiv.org/abs/1705.07874

[7] E. Mathieu, H. Ritchie, and L. Rodés-Guirao. 2020. *Coronavirus pandemic (COVID-19).* https://ourworldindata.org/coronavirus
[8] Mattiuzzi and Giuseppe L. [n. d.]. Analysis of Google Searches for COVID-19 and its symptoms for predicting disease epidemiology in the United States. *National Library of Medicine* ([n. d.]). https://doi.org/10.23750/abm.v92i1.11070
[9] S. Müller, M. Balmer, A. Neumann, and K. Nagel. 2020. Mobility traces and spreading of COVID-19. *medRxiv* (2020).
[10] A. Rovetta. 2021. Google trends as a predictive tool for covid-19 vaccinations in Italy: Retrospective Infodemiological Analysis (preprint). *National Library of Medicine* (2021). https://doi.org/10.2196/preprints.35356
[11] R. Salgotra, M. Gandomi, and A.H. Gandomi. 2020. Time series analysis and forecast of the COVID-19 pandemic in India using Genetic Programming. *Chaos, Solitons Fractals* 138 (2020), 109945.
[12] N.S. Sattar and S. Arifuzzaman. 2021. Covid-19 vaccination awareness and aftermath: Public sentiment analysis on Twitter data and vaccinated population prediction in the USA. *Applied Sciences* 11, 13 (2021), 6128.
[13] A. Shaham, G. Chodick, V. Shalev, and D. Yamin. 2019. Personal and social patterns predict influenza vaccination decision. *BMC Public Health* 2 (2019).
[14] R. Shortall, N. Moutera, and B. Van Weeb. 2022. COVID-19 passenger transport measures and their impacts. *Transport Reviews* 42, 4 (2022), 441–466. https://doi.org/10.1080/01441647.2021.1976307
[15] N. Sigalo, N. Awasthi, S.M. Abrar, and V. Frias-Martinez. 2023. Using covid-19 vaccine attitudes on Twitter to improve vaccine uptake forecast models in the United States: Infodemiology study of Tweets. *JMIR Infodemiology* 3 (2023).
[16] O. Wahltinez et al. 2020. COVID-19 Open-Data: curating a fine-grained, global-scale data repository for SARS-CoV-2. (2020). https://goo.gle/covid-19-open-data Work in progress.

8 APPENDIX

Code can be found here: https://www.kaggle.com/code/amish30/vaccination-prediction

A ADDITIONAL FIGURES

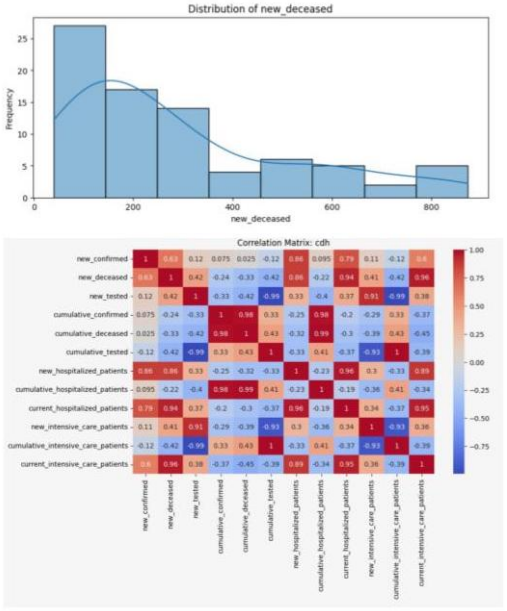

Figure 12: Example of EDA Analysis regarding Distribution and Correlation