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

Presented by Kartik Narang and Amish Saini

Introduction: Why estimate Vaccination Uptake?

Problem: Vaccination Uptake is hard to forecast and successful models are not interpretable.

Goal: Forecast Vaccination Uptake using interpretable techniques and modeling in order to inform health officials on how they can increase vaccination uptake.

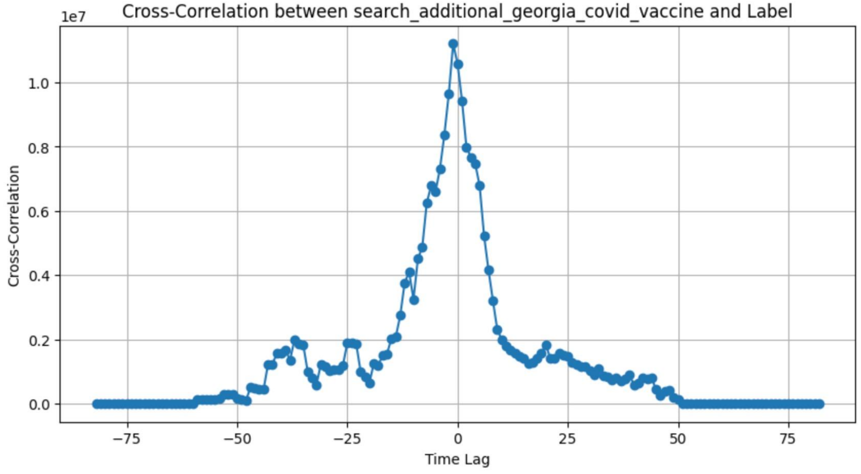
Why?

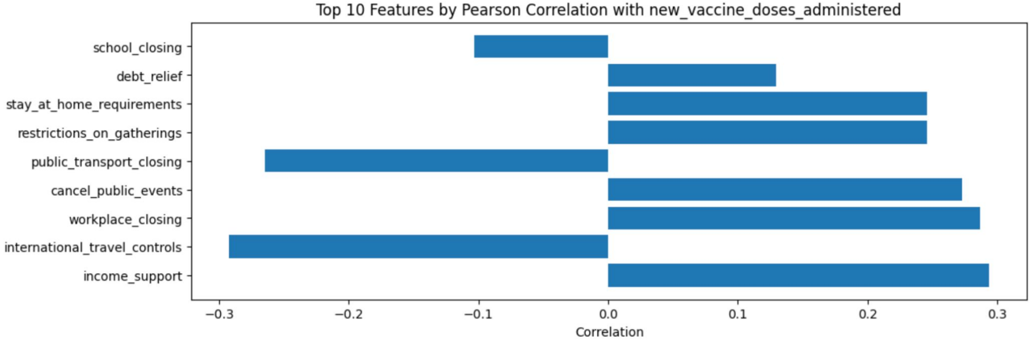
- Global Impact of COVID-19 and pandemics:
  - Worldwide impact of the pandemic, resulting in regional disparities.
  - Emergence of disparities highlights the need for effective predictive modeling.
- Challenges in Vaccination Rates:
  - Periods of stagnation in US vaccination rates despite widespread availability.
  - Identifying factors impacting vaccination uptake crucial for informed decision-making.
- Demystifying Individual Factors that lead to vaccination decisions

Dataset

Google COVID-19 Open Data Repository

- Daily information about topics related to COVID-19
- Accumulation of various datasets
  - Vaccinations
  - Search Trends
  - Epidemiology
  - Mobility
  - Weather
  - Miscellaneous (Economic, Local and Governmental Policies, etc.)
- 574 features
- 80 weeks: January 2021 to September 2022
- Cross Correlation: understand temporal relationships by determining time lag
  - (-1) means 1 day lead time in comparison to vaccination uptake
- Pearson Correlation
  - 1 of 3 total that we run

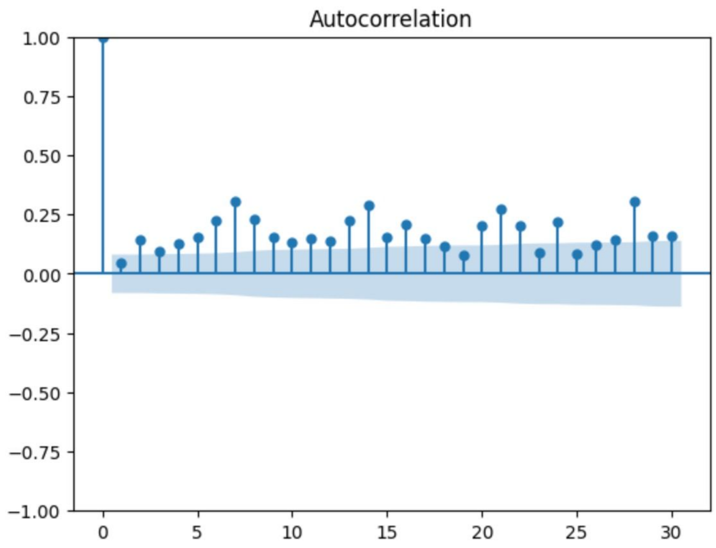


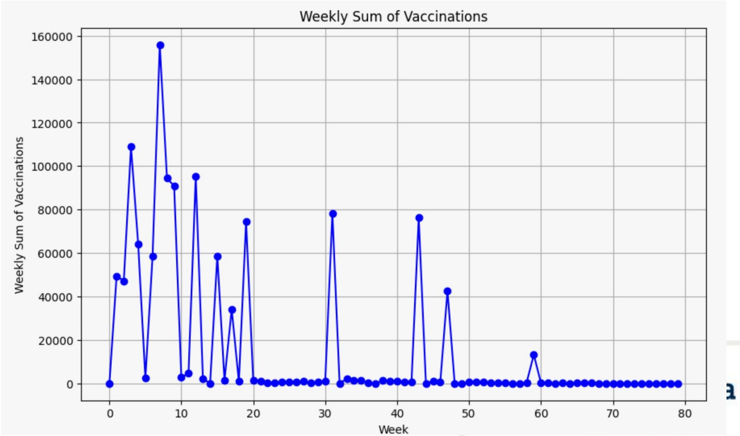


Dataset

Google COVID-19 Open Data Repository

- Previous Label: Number of Daily Vaccines
- Autocorrelation plot values - updated at the end of each week
- Spikes at 7, 14, 21, 28
- New Label: Number of weekly vaccinations
- Weekly Sum of Vaccinations: Training Data had heavy vaccination spikes in the beginning





Tree-Based Regressors

XGBoost

- Extreme Gradient Boosting
- Number of decision trees, each tree trained on a subset of data
- Combine predictions from each tree to compose final prediction
- Weak learners (regression trees) combined together to make one strong learner
- Gradient Descent to minimize loss

AdaBoost

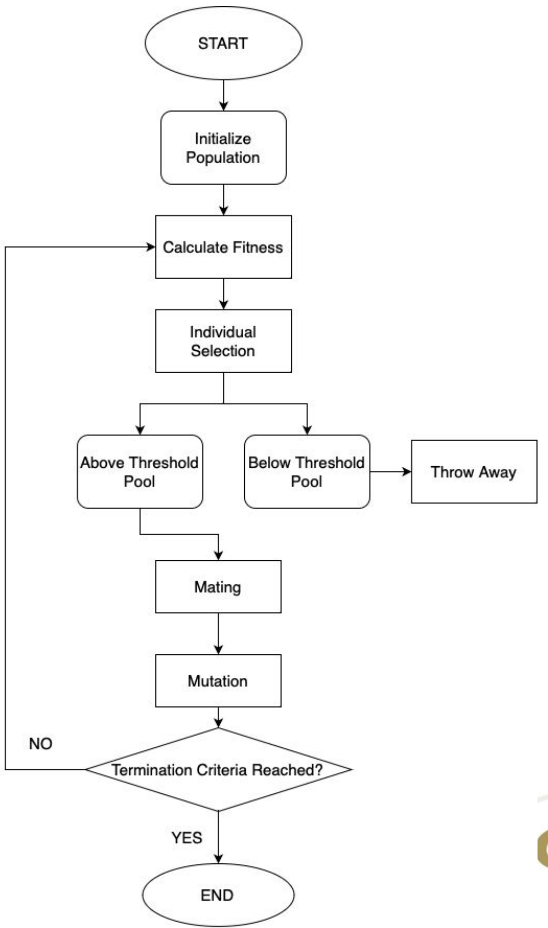
- Adaptive Boosting
- Less prone to overfitting since input parameters are not jointly optimized
- Adjusts weights to focus on misclassified instances (larger weight to incorrectly classified instances)

Random Forests

- Bagging Approach
- Grows out multiple decision trees and merge them together for final prediction
- Trees constructed using random subset of training data and random subset of features at each split (not sequentially)
- Randomness: reduces overfitting, better generalization

Genetic Programming - Symbolic Regression

- Symbolic Regression:
  - Subset of GP utilizing symbolic regression.
  - Searches mathematical formula space for optimal predictors.
- EA Framework:
  - Baseline algorithm employs Evolutionary Algorithm.
  - Population starts with randomized expressions and evolves over generations.
- Fitness Evaluation:
  - Fitness based on performance metrics: RMSE, MAE, R^2.
- Mating and Mutation:
  - Mating introduces gene crossover with varied methods.
  - Mutation adds diversity with different methods and probabilities
- DEAP Library Implementation



Georgia Tech

Interpretability

Tree-Based Regressors

- SHAP (SHapely Additive exPlanations)
  - Game theoretic approach to explain output of ML models
- Uses Shapley values to assign credit for a model's prediction to each feature.
- Tree explainer: Assess feature importance for ensemble model of XGBoost, AdaBoost, and Random Forest
- Sheds light on feature interactions and their influence on model predictions
- For our model's forecast, we can see exactly which feature pushed or contributed to that value

Genetic Programming

- Examine presence and frequency of features in the expression
- If feature shows up multiple times across different runs of GP → high importance
- Impact of feature removal on model performance
- Relative contribution of each feature to the overall fitness of the expression

Experiment 1 & Results - Tree-Based Regressors

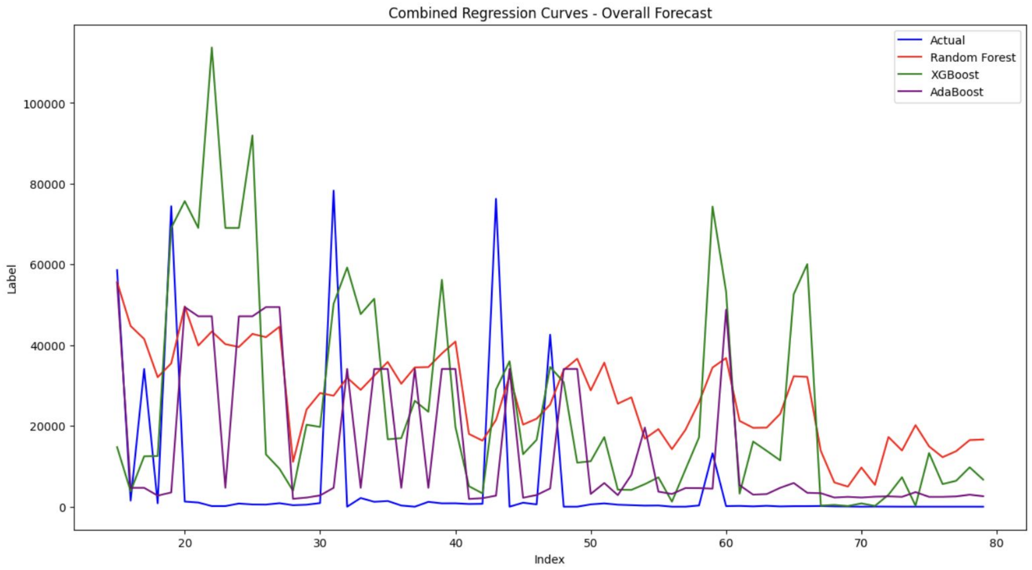
Experiment

- Time Series Split
  - Expanding window of various train and test sizes
- Average Metrics over all Forecasts
- Fine-tune Models for best performance
  - Parameters
  - Various Feature Selections

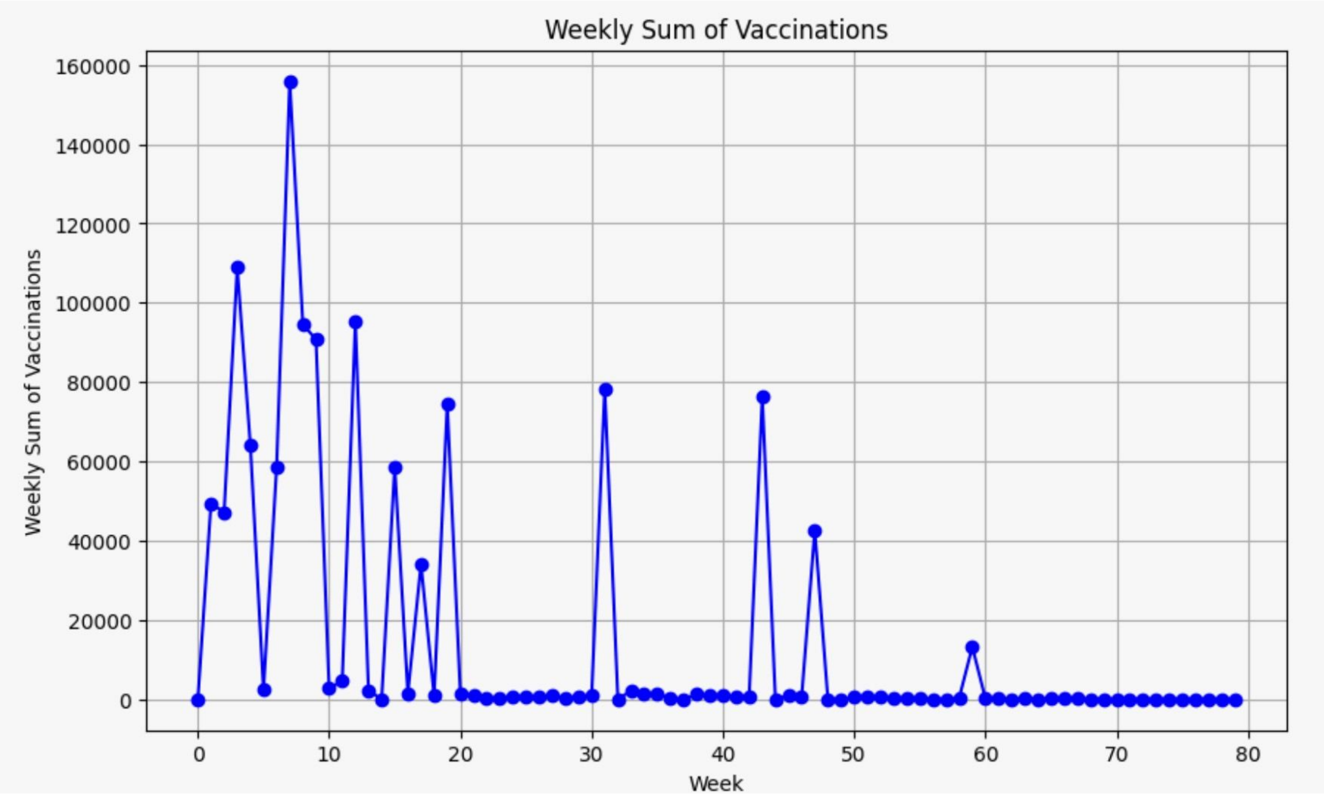
Results

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



Major Issue



$\underbrace{\hspace{4cm}}_{\text{Train}}$ $\underbrace{\hspace{9cm}}_{\text{Forecast}}$

Experiment 2 & Results - Tree-Based Regressors

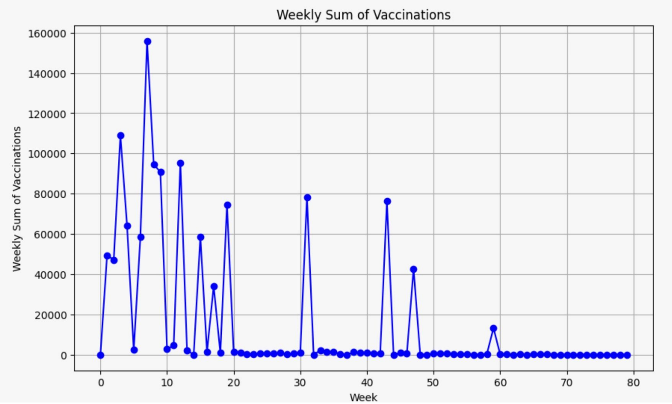
Experiment

- Train on initial spikes
- Forecast Future Spikes
- Fine-tune Models for best performance
  - Parameters
  - Various Feature Selections

Results

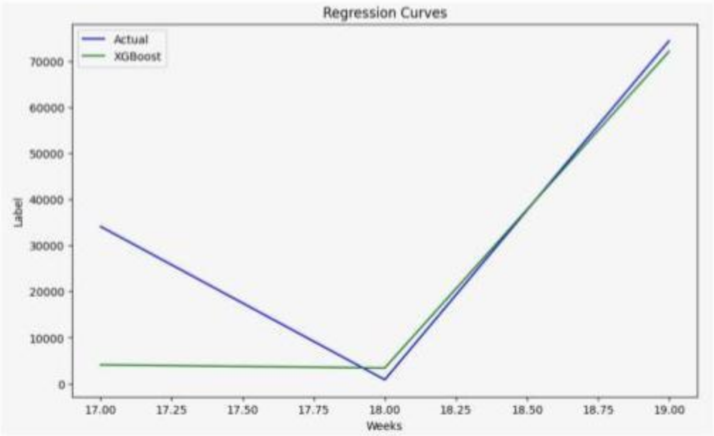
Table 2: Spike Forecast Tree-Based Regressor Metrics

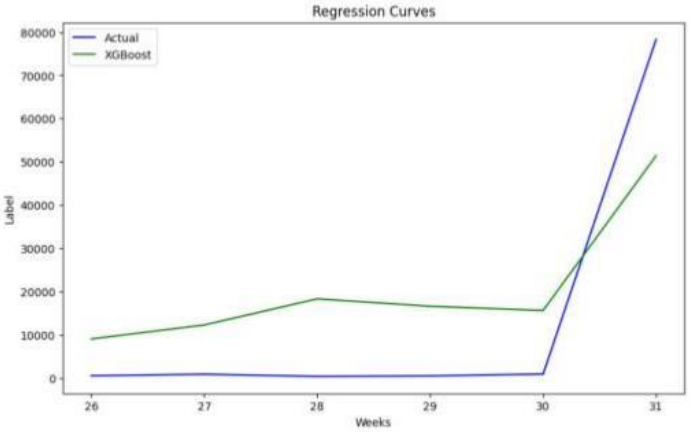
Model Type	Spike Start Week	Norm. RMSE	Norm. MAE	R-Squared
XGBoost	18	0.2373	0.1580	0.6631
XGBoost	30	0.2175	0.2044	0.6572
AdaBoost	42	0.2291	0.1153	0.6204

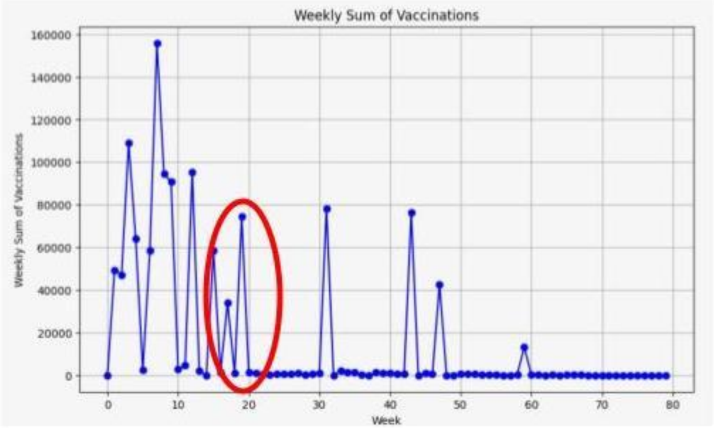


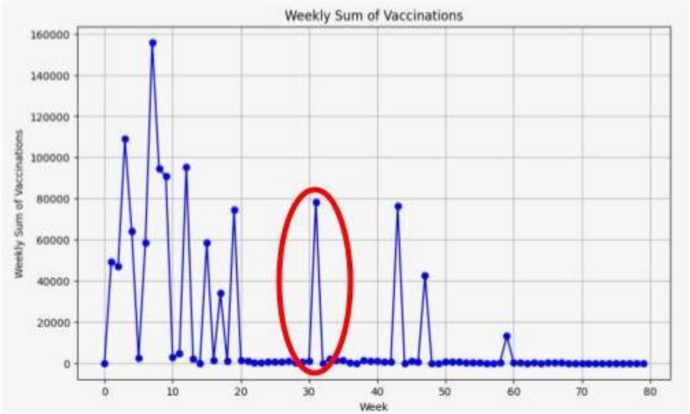
Georgia Tech

Results - Tree-Based Regressors Continued



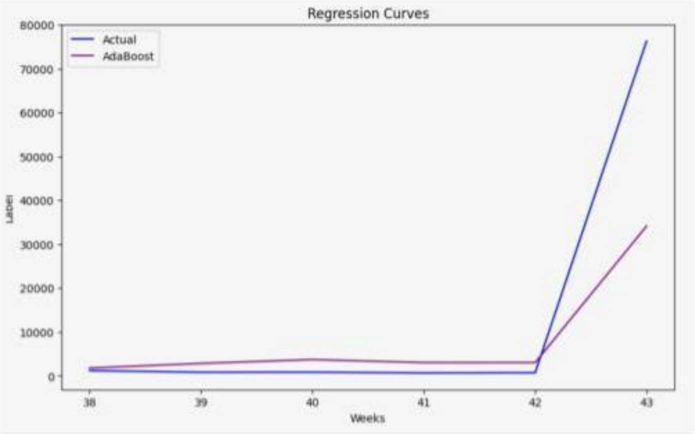


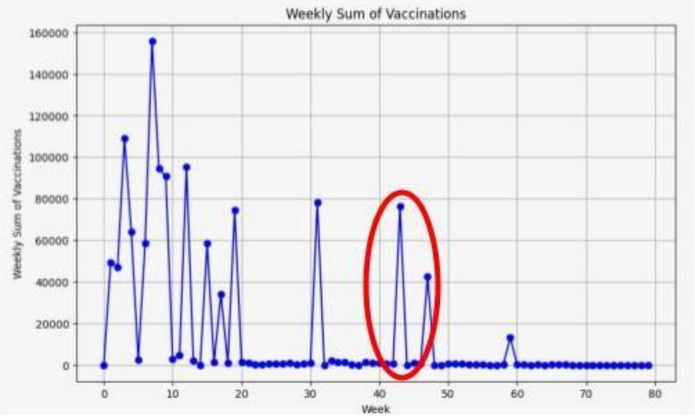




Georgia Tech

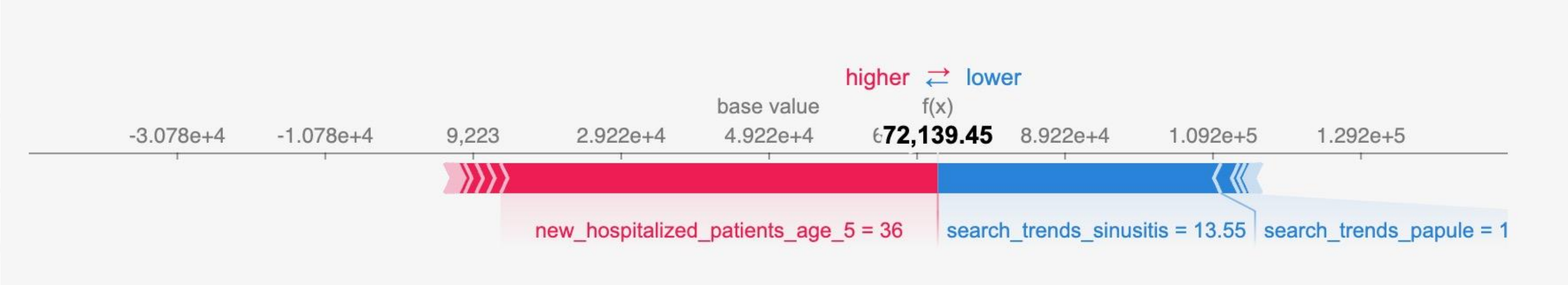
Results - Tree-Based Regressors Continued

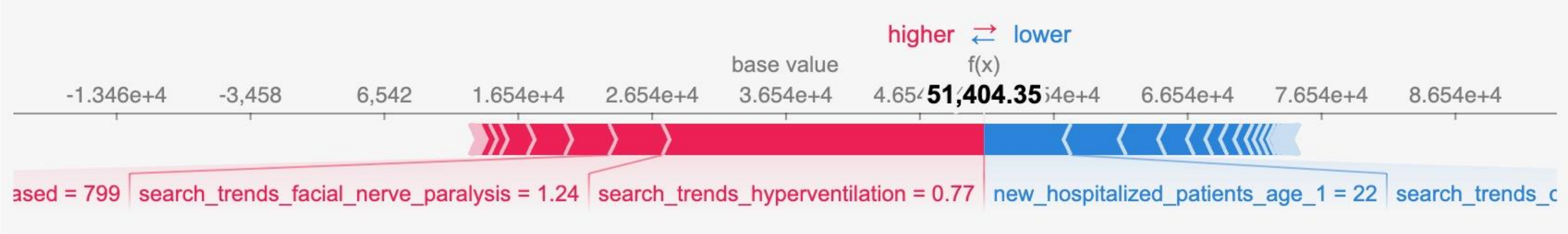




Georgia Tech

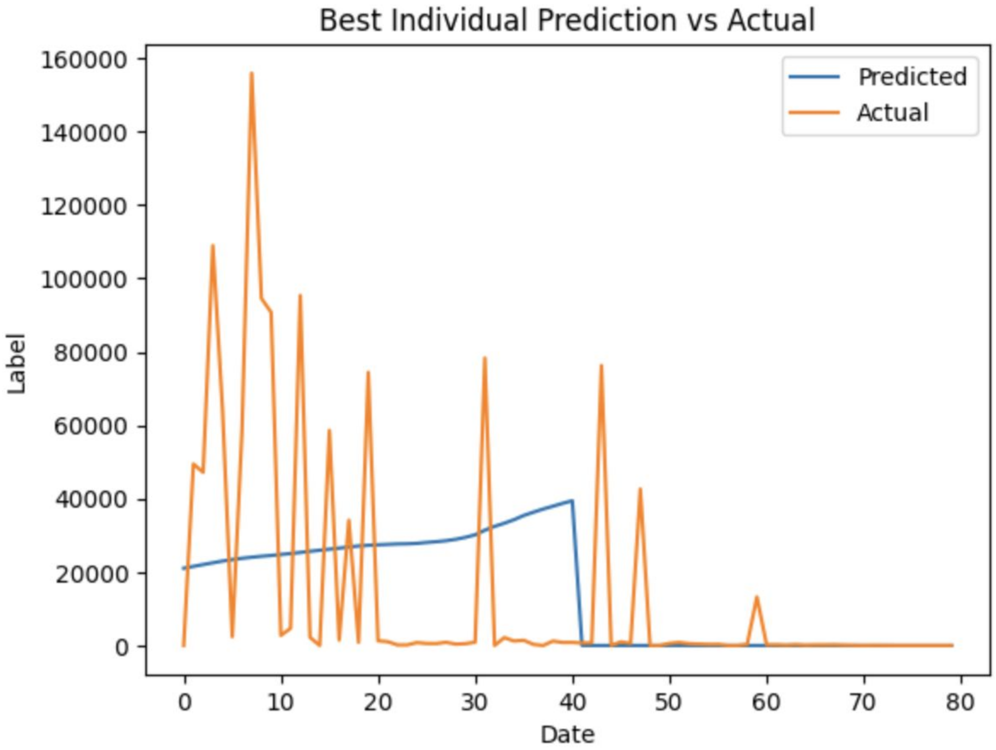
Interpreting Tree-Based Results - SHAP





Genetic Programming Experiment & Results

- Generations = 500
- Population = 1000
- Mating Probability = 0.5
- Mutation Probability = 0.3
- Operators:
    - Add
    - Subtract
    - Cos
    - Sin
    - Multiply
    - Divide
    - Exp
- Not enough computational power!



Georgia Tech

Conclusion and Future Work

Predicting Vaccination Uptake is a difficult task

Previous Research only successful using models like ARIMA
- Not Interpretable
- Not very useful in helping officials make decisions

Tree-Based Regressors
- Forecast peaks with sufficient accuracy
- Interpret feature impacts that led to model's predictions (Validity Issue)

GP
- Failed due to difficulty of fitting entire curve
- Computational Complexity Bottleneck

Next Steps
- Less Scarce Data, allowing for higher confidence in predictions and interpretation
- Tree-Based Regressors to forecast on a disease that has data for much longer time period (Flu) to get more confident predictions and interpretations.

Georgia Tech