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

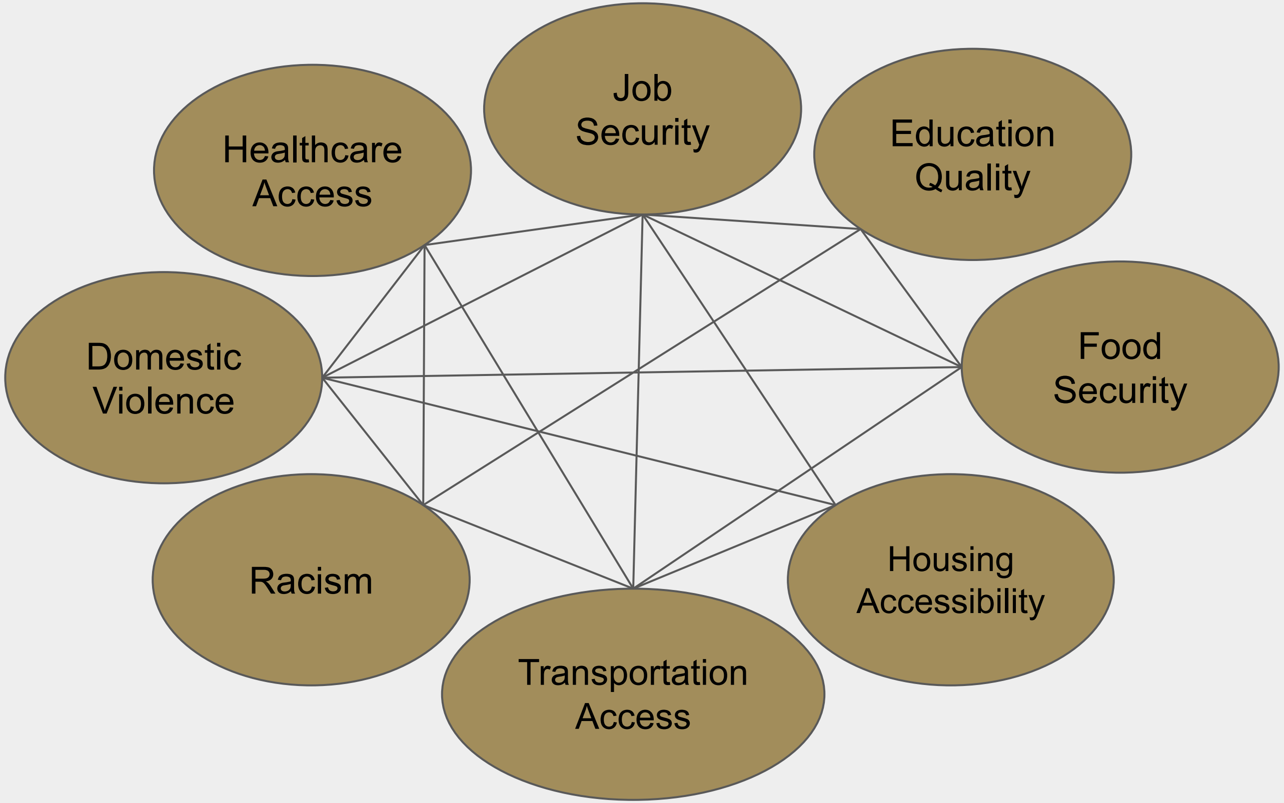
Andy Chea, Saideep Narendrula, Rachel Calder

CSE8803 Final Presentation

Overview

- Introduction and Motivation
- Our Approaches
- Data Description
- Results
- Evaluation of Results
- Conclusions and Future Directions

Global Pandemic Impacts



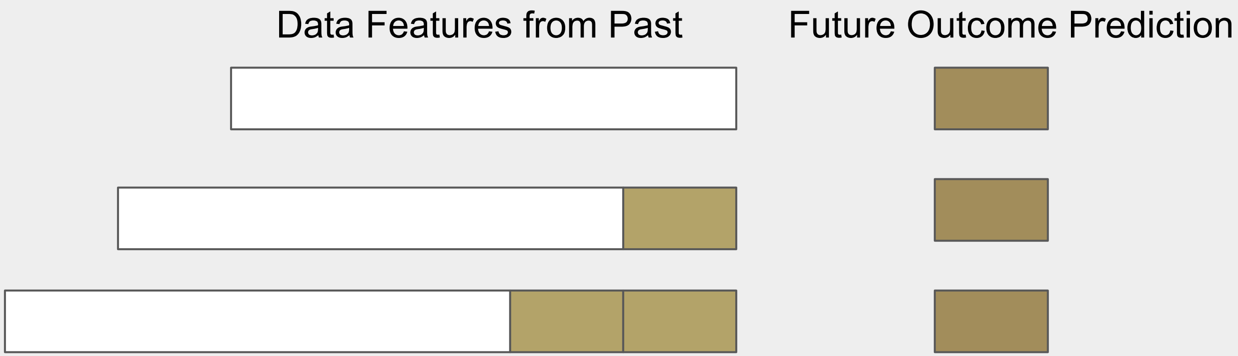
CITE [1]

Impacts of Forecasting

- Strategic Intervention
- Decrease in cases and deaths
- Resources saved
- Improvement of personal social determinants of health outcomes

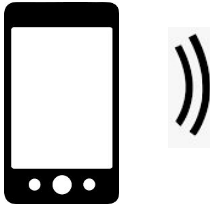
CITE [2]

Forecasting Techniques



- What features are sufficient for forecasting?
- Is it better to fewer features for a longer time period?

Our Data



1. Restaurant Visit
2. Bars Visit
3. Survey for COVID Indicators

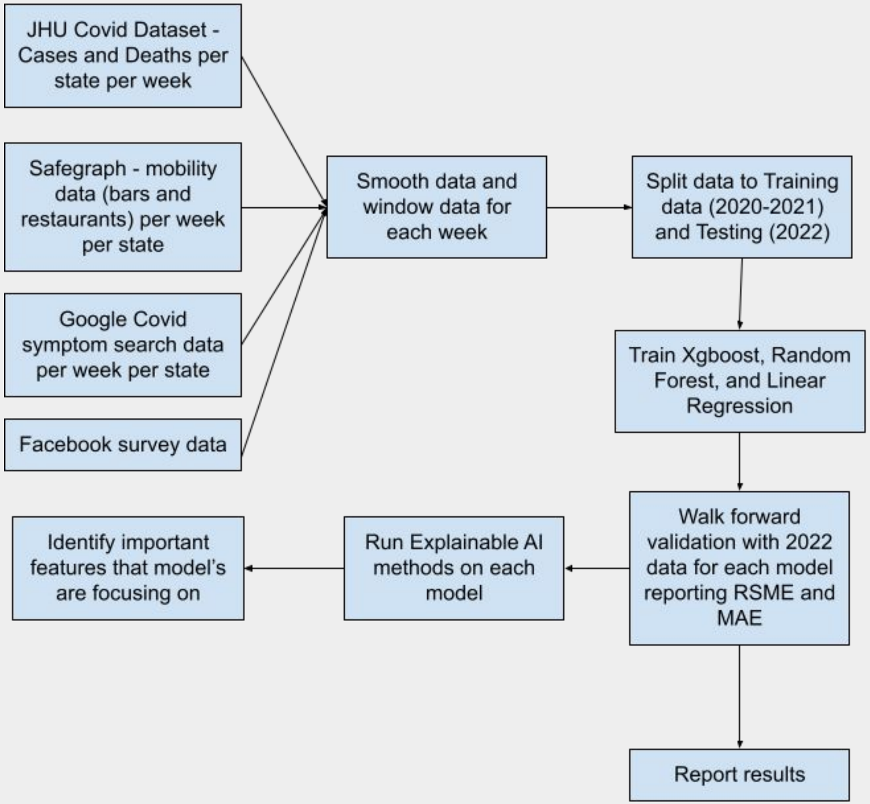
4. Cases
5. Deaths

CMU Delphi API

Fever
Major depressive disorder
Food Intolerance
Hepatitis
Fatigue
Migraine
Hives
Runny Nose
Chills
Eye strain
Gingival recession
Hot flash
….+410

Google Symptom Search Trends

Our Strategy



Forecasting Models

Explore ML, statistical, and deep models

1. **Random Forest Regressor (ML) -** ensemble model made up of bagged decision trees. Both branch points and datasets have random aspects.
2. **XGBoost Regressor (ML) -** Gradient boosting
3. **Linear Regression (stat)** - Fits a line that minimizes the sum of the squared differences between prediction and ground truth.
4. **Long Short-Term Memory / CAMul (deep)** - multi-layered analyses
   a. Not explainable with SHAP
   b. Attempted rewriting the SHAP package for LSTM and CAMul

Daily Walk-Forward Validation

- Scale the data since they're from different datasets
- Split it into input (FB survey, symptom search, mobility) and output variables (cases or deaths)
- Offset the features from response so that inputs from one day are fitted to output for the next.
- Train on the first 650 days initially, then walk-forward validate for 133 predictions to reach the day count of 788 days per chosen state.
- Record the current prediction then retrain the model with the ground truth value added to the training set.
- Assess the performance using MAE and RMSE

Evaluation Metrics

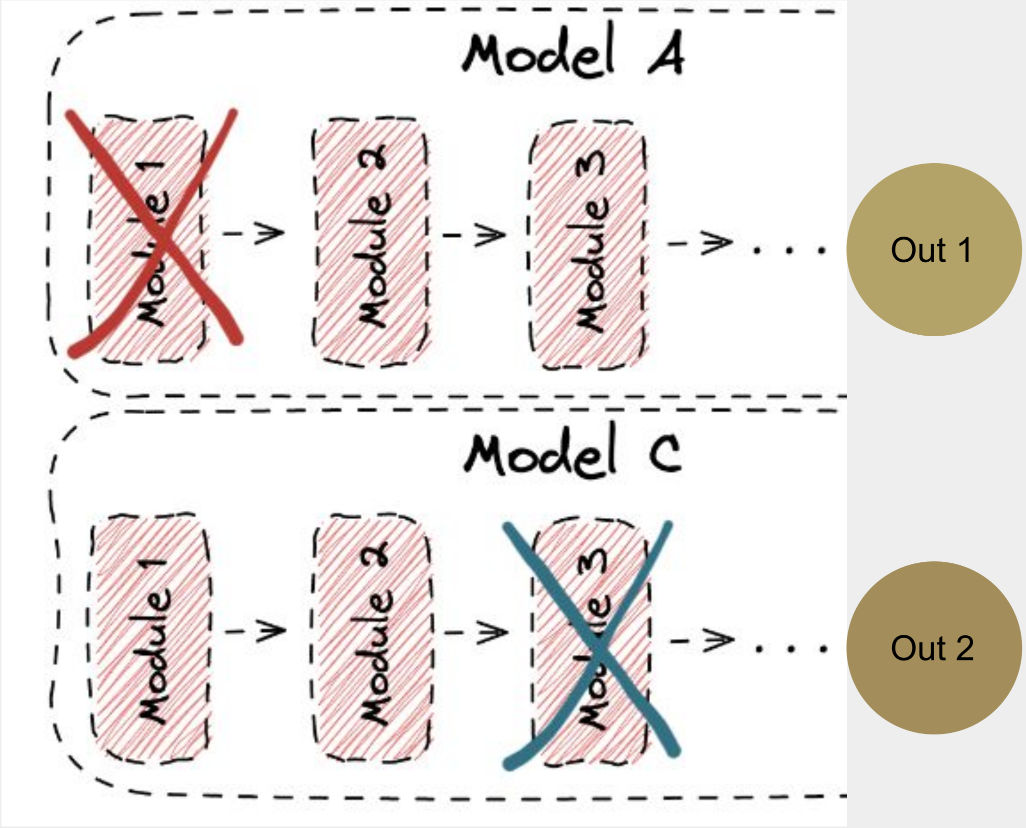
MAE was chosen as an intuitive measure of error. Monotonically increases

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

RMSE was chosen to evaluate models with large deviations heavily penalized.

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

Explainable AI: Ablations



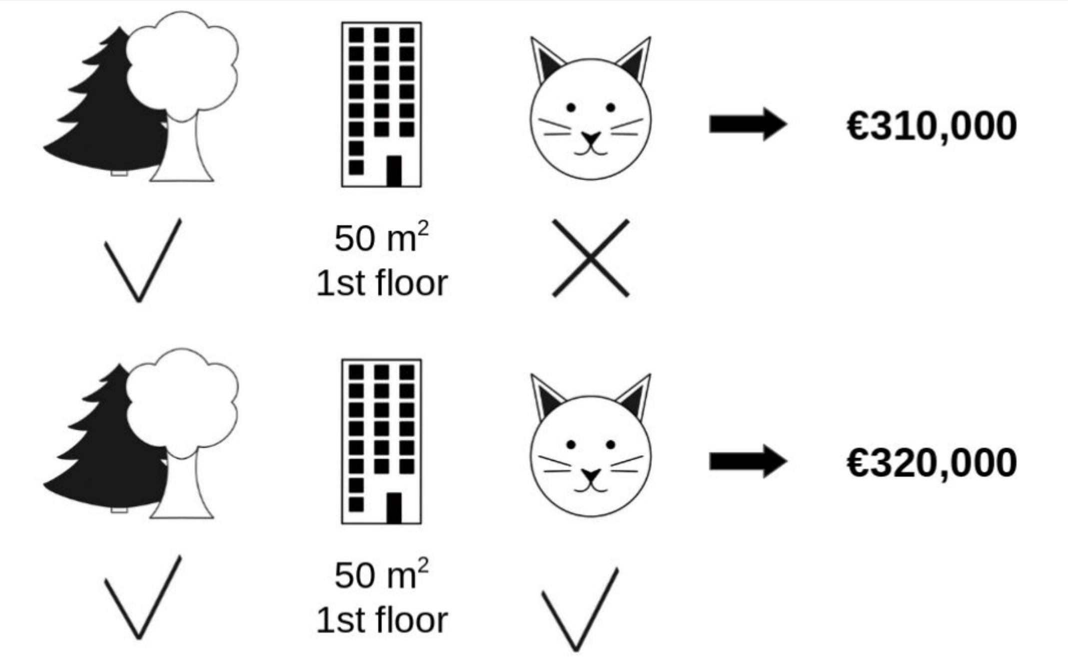
- From biology, surgically removing certain portions.
- Drop out certain features and then retrain the model with the truncated set, and determine effect on output.
- Add 25 at a time in an order defined by each feature's Pearson Correlation Coefficient.

Explainable AI: SHAPley Values

From game theory, figures out how much each player contributes to a certain payout.

Obtain a marginal value in all possible coalitions when a given feature is added.

The Shapley value is the <u>average</u> contribution of a feature value to the prediction in different coalitions.



Forecasting: Model Results

Xgboost was the best performing model over the entire datatest with Random Forest performing very similarly

RSME for case forecasting was lower in all states except for California

As expected, machine learning models were better forecasters than Linear Regression

	Xgboost		Random Forest		Linear Regression	
	Cases	**Deaths**	**Cases**	**Deaths**	**Cases**	**Deaths**
MAE	0.035	0.279	0.045	0.277	0.909	1.338
RSME	0.119	0.368	0.182	0.367	3.964	6.55

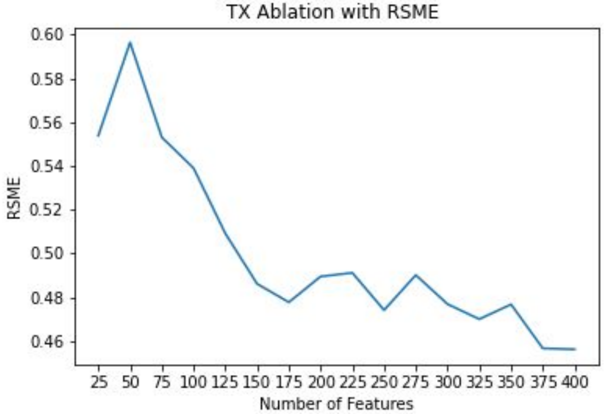
Explainable AI: Ablations Results

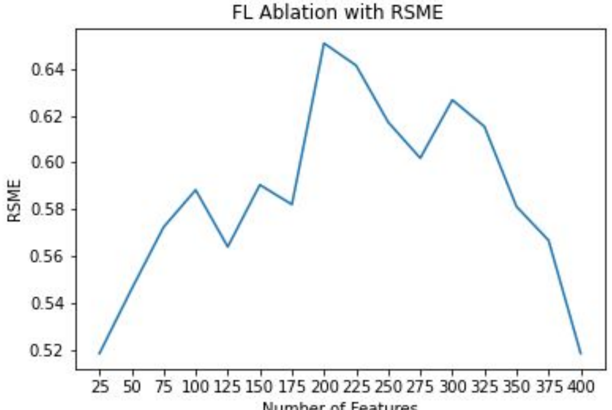
Ablations showed that with the increase of features, the performance increased

Texas, California, Massachusetts, and Georgia all had similar results

Florida performance suffered with the addition of features

- Important features for Florida were negatively correlated to the number of deaths

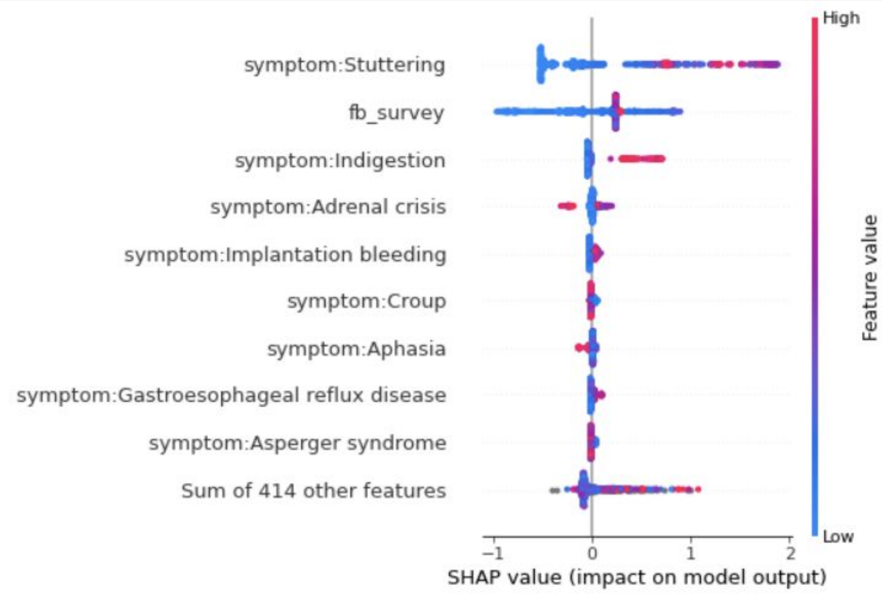




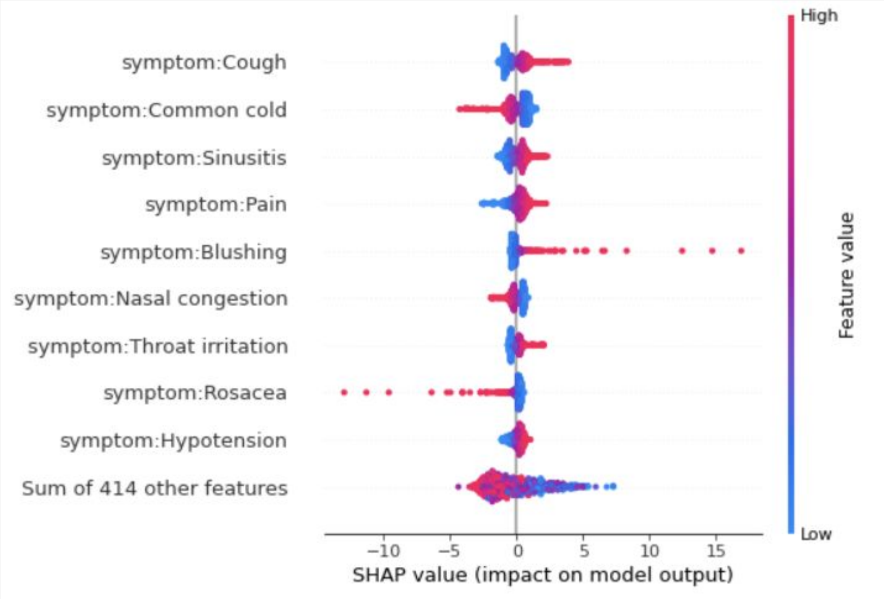
Explainable AI: SHAPley Analysis

XGBoost SHAP Beeswarm plots

Massachusetts



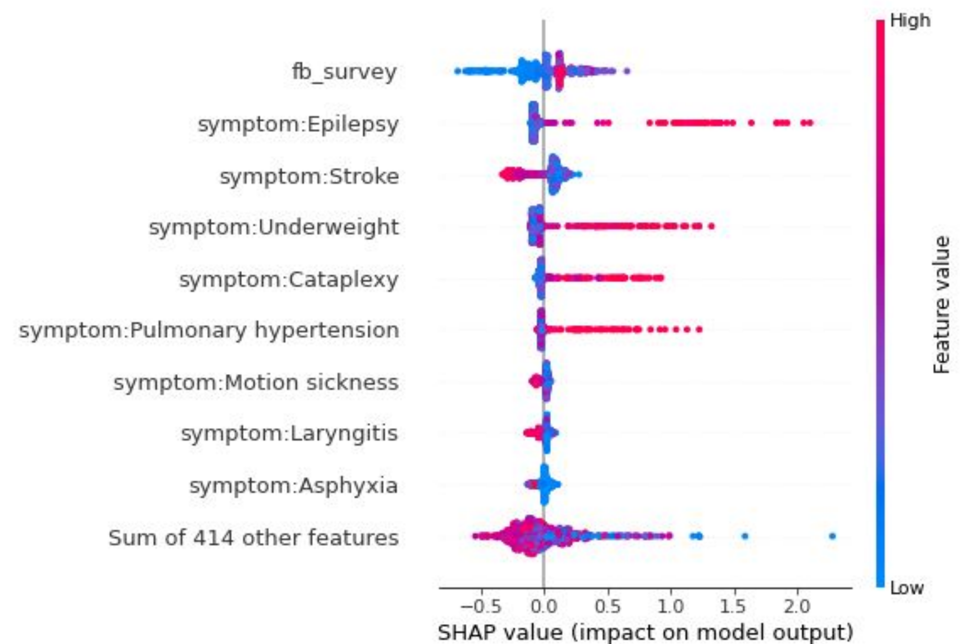
California



Explainable AI: SHAPley Analysis

Random Forest SHAP Beeswarm plots

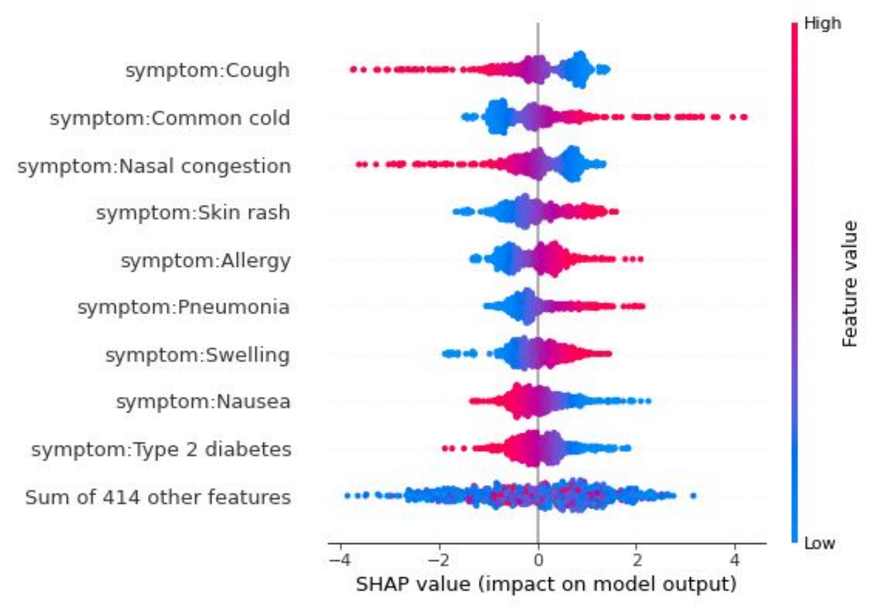
California



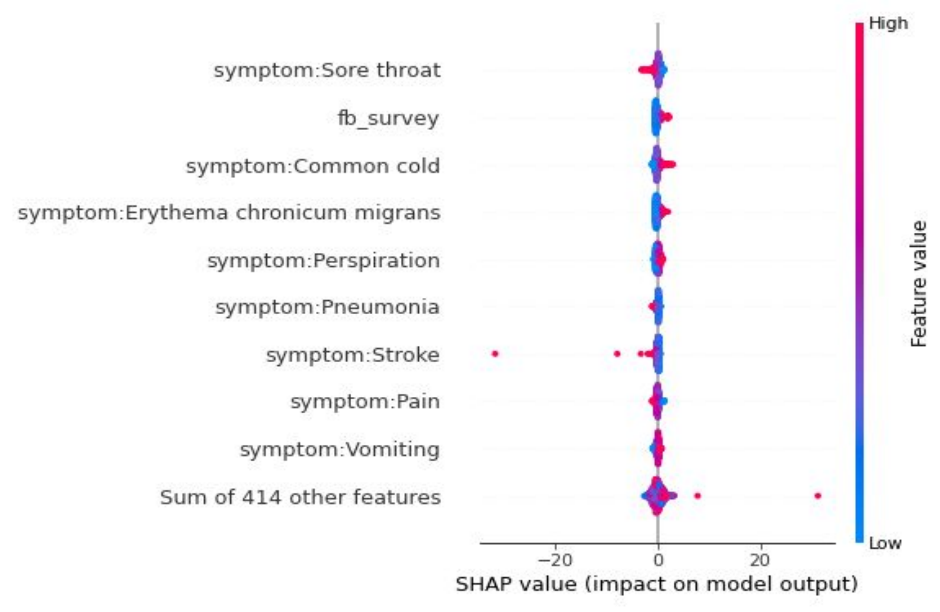
Explainable AI: SHAPley Analysis

Linear Regression SHAP Beeswarm plots

Texas



Florida



Discussion and Future Work

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

Future work should include combining more models for informative signals through finding novel ways to harmonize SHAP with them.

References

[1] Heidi Green, Ritin Fernandez, and Catherine MacPhail. 2021. The social determinants of health and health outcomes among adults during the COVID-19 pandemic: A systematic review. *Public Health Nursing* 38, 6 (2021), 942–952. DOI:http://dx.doi.org/10.1111/phn.12959

[2] Konstantinos Nikolopoulos, Sushil Punia, Andreas Schäfers, Christos Tsinopoulos, and Chrysovalantis Vasilakis. 2021. Forecasting and planning during a pandemic: Covid-19 growth rates, supply chain disruptions, and Governmental Decisions. *European Journal of Operational Research* 290, 1 (2021), 99–115. DOI:http://dx.doi.org/10.1016/j.ejor.2020.08.001

[3] 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

[4] 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

[5] 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