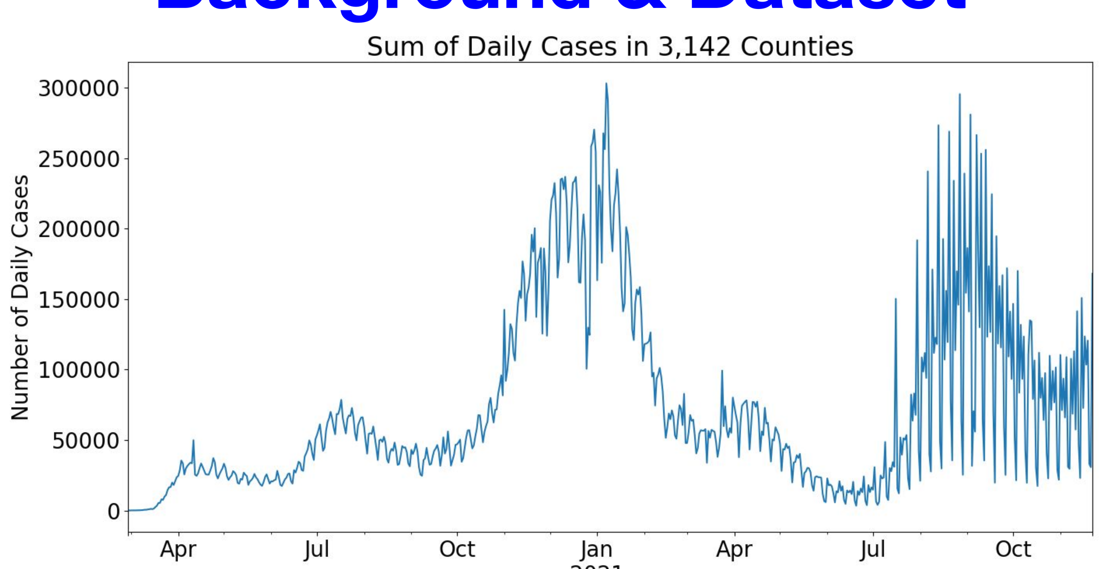
Understanding County Level Covid-19 Infection and Feature Sensitivity with Deep Learning

Andrej Erkelens, Nick Daniello, Md Khairul Islam, Di Zhu, Judy Fox
University of Virginia

Abstract

In this work we study how deep learning models forecast COVID-19 using the state-of-the-art time-series forecasting model called Temporal Fusion Transformer (TFT) from Google [1]. Then we use sensitivity analysis with Morris Method [2] to see how sensitive the models is with respect to the feature. We have collected a dataset with 3142 counties, 2 target variables (cases and deaths), and a number of static and dynamic features. Using the processed data, we show our model performance on the 130 most populous counties in the US. Being able to model the disease at a county level allows administrations to make decisions based on what the disease will look like in the near future.

Background & Dataset



The dataset used in this work is collected from multiple sources. All are at county level. The ground truth daily covid cases is collected from USAFacts (shown in the plot above). Static feature are listed in the table below:

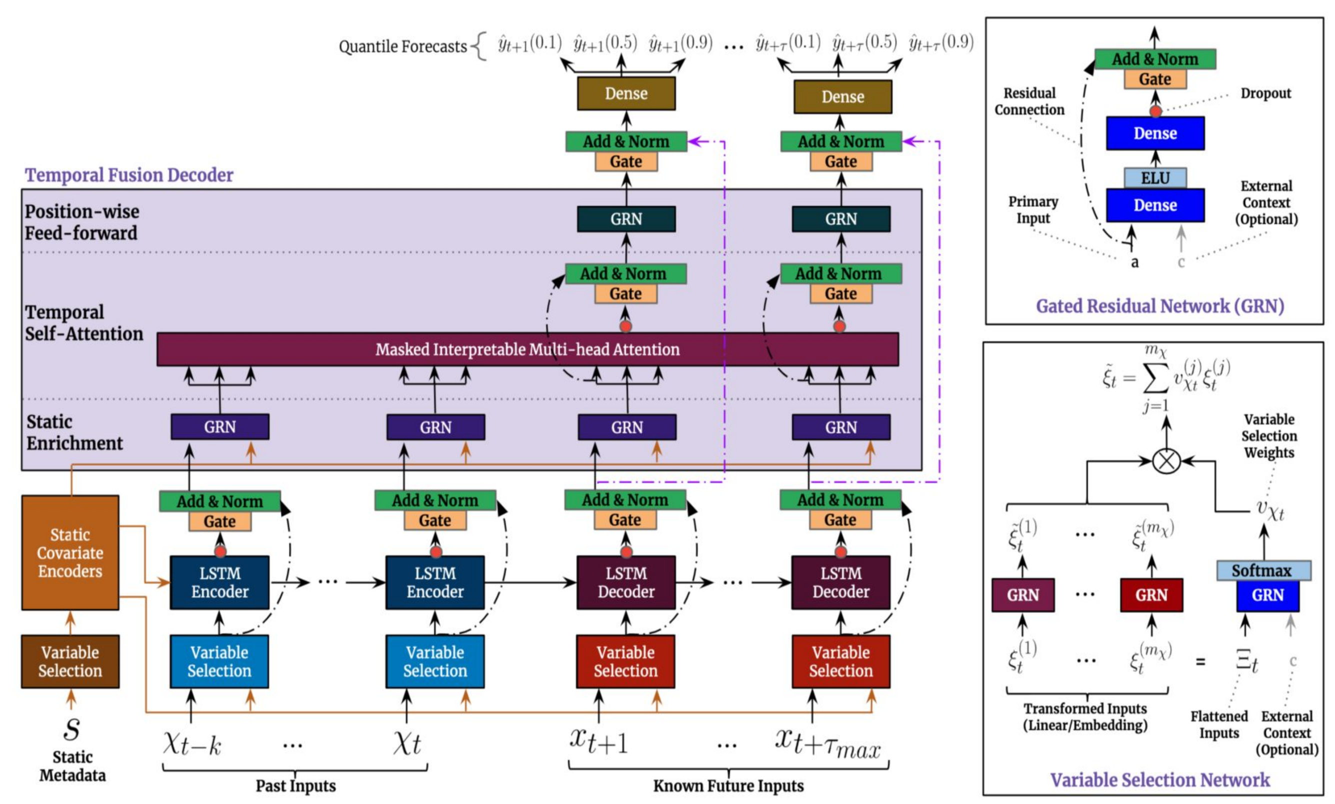
Feature	Definition	Source
Age Distribution	% age 65 and over	CDC SVI
Air Pollution	PM2.5	County Health Ranking
Health Disparities	Socioeconomic Status and % Uninsured	CDC SVI

Following is the table for dynamic features :

Feature	Definition	Source
Disease Spread	Fraction of total cases from the last 14 days	NIH
Transmissible Cases	Cases from the last 14 days divided by Population	NIH
Vaccination	$ of population with series complete	CDC
Social Distancing	based on mobility	Unacast. NIH PVI dashboard

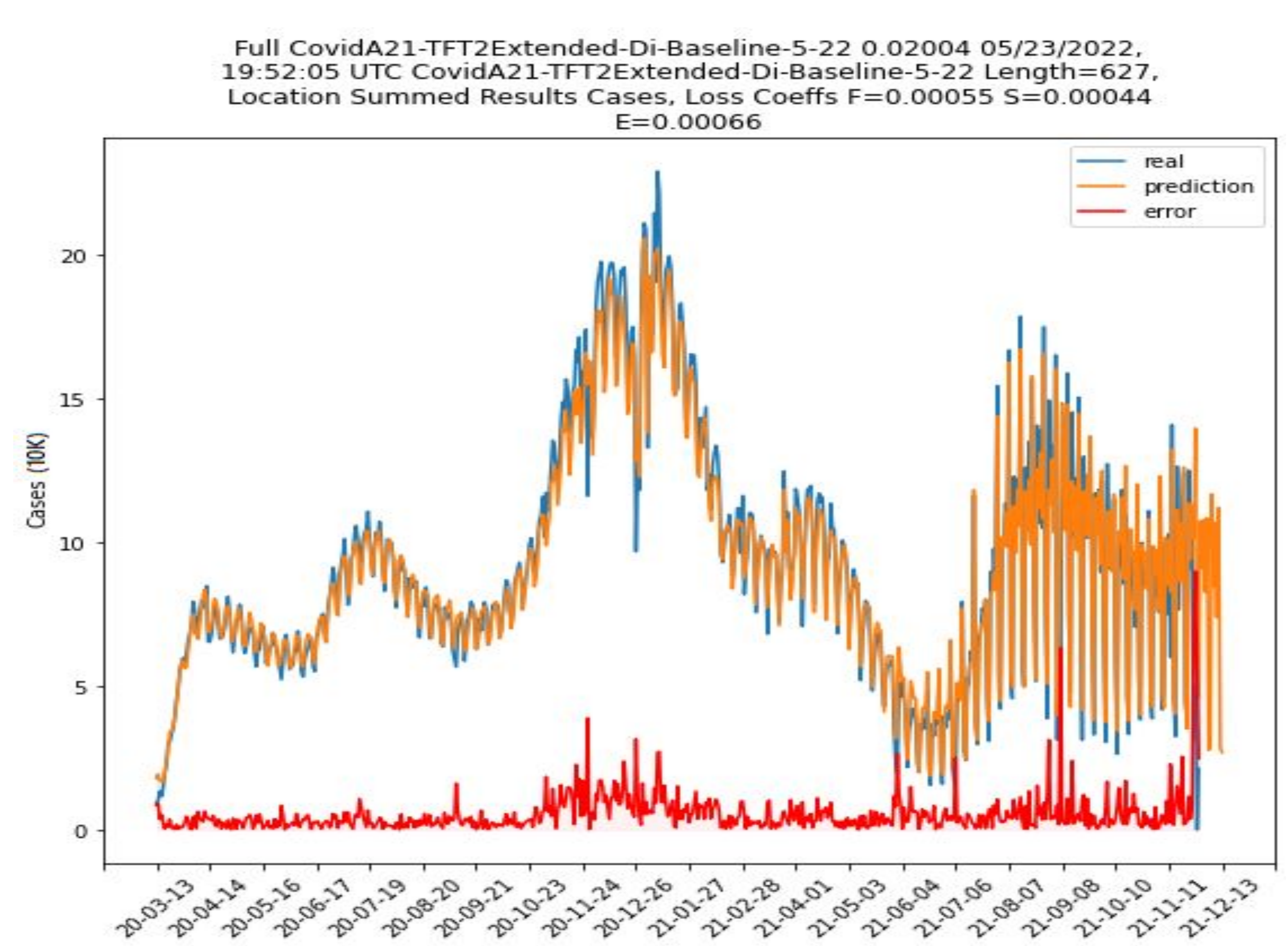
Model

We used the TFT model [1] for the disease cases and deaths. TFT takes static metadata, time-varying past inputs and time varying a priori known future inputs. Variable Selection is used for judicious selection of the most salient features based on the input. Gated Residual Network blocks enable efficient information flow with skip connections and gating layers. Time-dependent processing is based on LSTMs for local processing, and multi-head attention for integrating information from any time step.



Prediction Results:

TFT shows significant results on the top 130 counties based on population. The performance metric we used was NNSE (Normalized Nash Sutcliffe Efficiency). The plot is summed across all time period. Bottom is visuals of cases prediction along with real data and error. Small error proves that TFT has good performance on cases prediction.

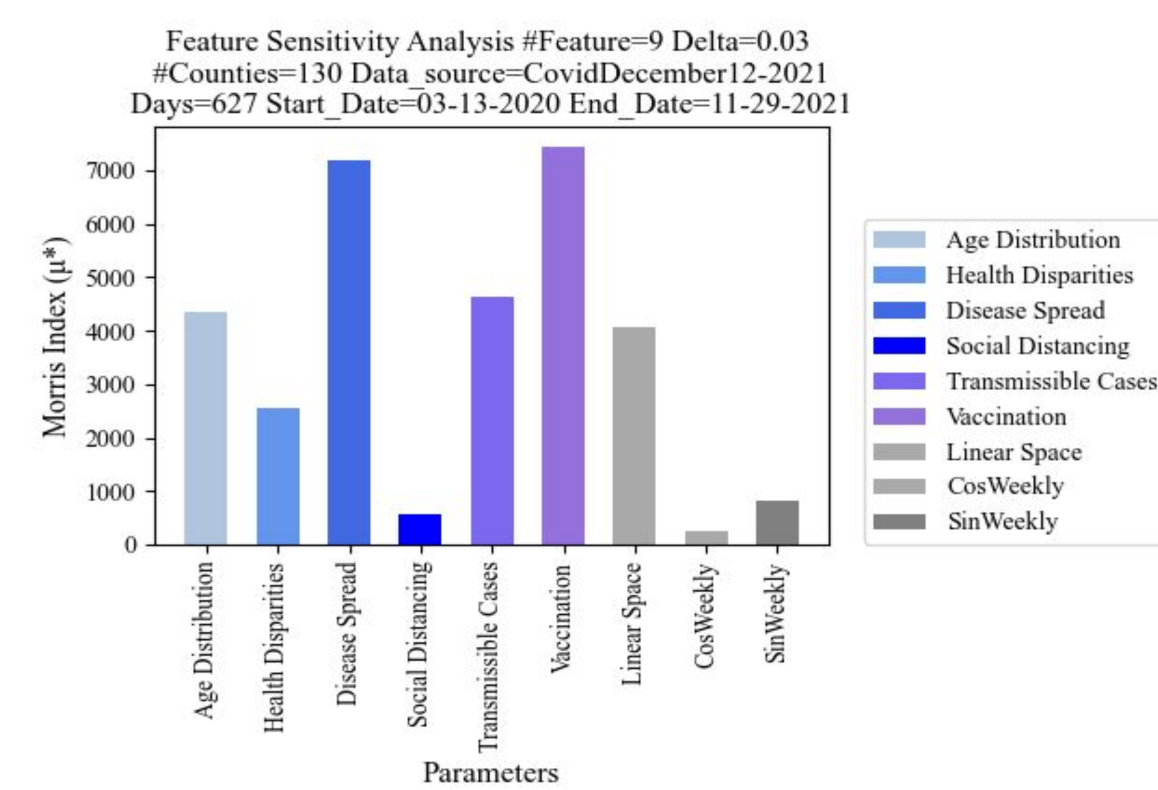


Feature Sensitivity

The sensitivity of a parameter x_i, is expressed by the Morris index μ defined as the ratio of the change in an output variable to the change in an input parameter.

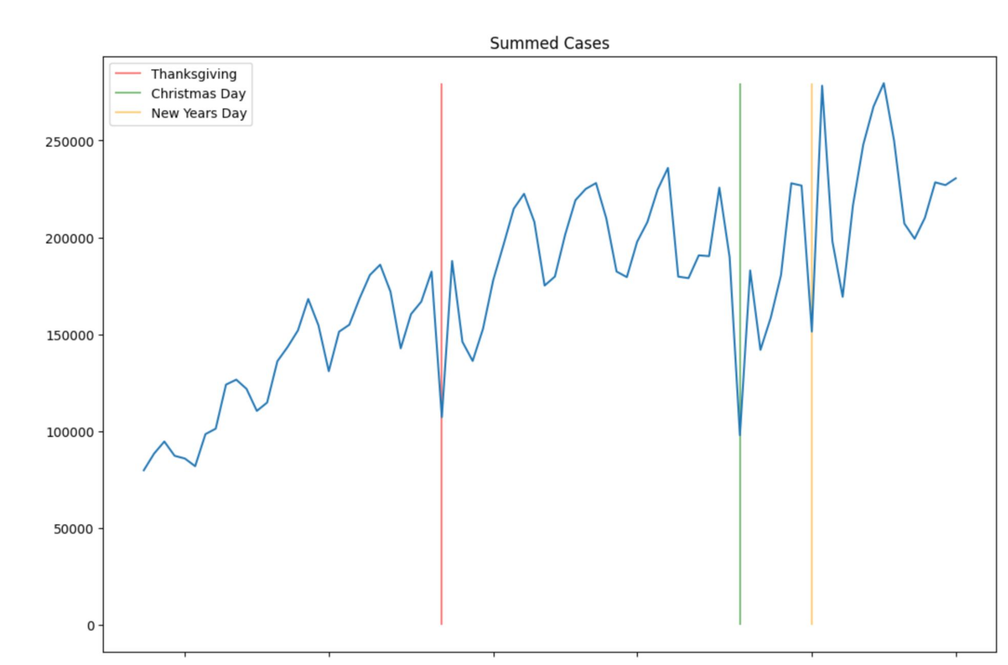
$$\mu_i^* = \left| \frac{y_j(x_1, x_2, \ldots, x_i + \Delta, x_{i+1}, \ldots, x_k) - y_j(\mathbf{X})}{\Delta} \right|$$

The Morris index is summed across all time periods and top 130 counties by population. Following is the plot for sensitivity analysis. Vaccination and Disease Spread are the most influential parameters.



Results: Trends

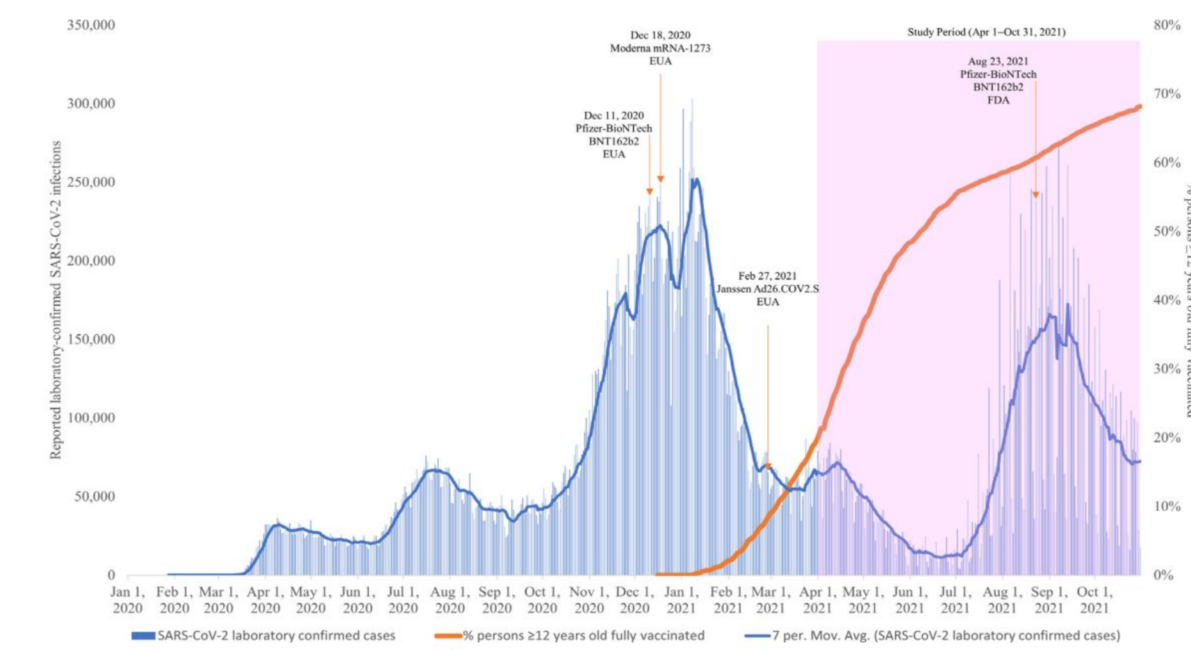
Additionally in the data, we notice some trends, for example the holidays (Christmas, NYE, Thanksgiving) experience extreme behavior as far as new cases go. TFT incorporates the ability to use these known events in the modeling. We can see this in aggregated plot below. Also note peaks and troughs in each week of data. This corresponds to Sundays (trough) and Fridays (peak) for cases data during this time period.



Conclusion

1) TFT works significantly well for learning different trends and events from the data.
2) The feature sensitivity study enable us to look into stratifying the modeling based on population groups, and how that effects administering covid responses to vaccination and other important features.

Related Work



[2] evaluated the association between county-level COVID-19 vaccine uptake and rates of COVID-19 cases and deaths in the United States The figure shows Daily laboratory-confirmed COVID-19 cases in the United States and percent of persons ≥12 years of age fully vaccinated, January 1, 2020 – October 31, 2021 .

Acknowledgement

This work is partially supported by NSF grant CCF-1918626 Expeditions: Collaborative Research: Global Pervasive Computational Epidemiology. We appreciate technical support from Intel Inc.

References

[1] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2020.

[2] County-level vaccination coverage and rates of COVID-19 cases and deaths in the United States: An ecological analysis.

[3] The repository for this work is available at **https://github.com/Data-ScienceHub/GPCE**

UVA DATA SCIENCE