Interpreting County Level COVID-19 Infection and Feature Sensitivity using Deep Learning Time Series Models

**Abstract**

* This project combines sensitivity analysis with heterogeneous time-series deep learning model.
* Forecast county-level COVID-19 infection using the Temporal Fusion Transformer (TFT)
* Sensitivity analysis extending Morris Method to see how sensitive the outputs are with respect to perturbation to our static and dynamic input features
* Goals of project:
  + Model can capture the detailed daily changes of temporal and spatial model behaviors and achieves high prediction performance compared to a PyTorch baseline.
  + By analyzing the Morris sensitivity indices and attention patterns, we decipher the meaning of feature importance with observational population and dynamic model changes
* Model the disease infection with a hybrid prediction and description accuracy measurement with Morris index at the county level

**Introduction**

* To effectively study county-level input features, we design a novel method to compute the Morris index but generalize it to multidimensional spatial and temporal variables.
* Using a self-attention-based Temporal Fusion Transformer (TFT) model, we can capture a complex mix and full range of static and dynamic covariates, known inputs, and other exogenous time series parameters.
* We perform individual feature importance evaluations to identify the most influential features for prediction and the sensitivity of infected cases.
* Contributions:
  + Introduce individual feature sensitivity to forecasting outputs with an extended Morris Method for multidimensional spatial and temporal data.
  + Model heterogeneous time-series prediction and analyze attention weights for insights on feature importance.
  + Stratify county-level population characteristics (Age and Industry segments) from socioeconomic and health data.

**Input data and features**

* Dataset for 3142 US counties
* Entries from 02-29-2020 to 05-17- 2022
* Select 9 observed features, static and dynamic, to predict cases and deaths.
* Focus on the following:
  + Age distribution
  + Health disparities
  + Disease spread
  + Social distancing
  + Transmissible cases features

**Background and Theoretical Foundation**

* The model inputs are passed through a Variable Selection Network (VSN) to select the most salient features and filter out noise.
  + Learning significant data points is done by leveraging local context with LSTM-based sequence to sequence layer.
* Sensitivity analysis can identify the importance of each model parameter in determining the outputs

Graphical user interface, text, application

Description automatically generated

* Morris method, a reliable and efficient sensitivity analysis method that defines the sensitivity of a model input as the ratio of the change in an output variable to the change in an input feature

**Experimental Setup**

* Implement the TFT model with both Tensorflow and PyTorch
* Runtimes:
  + The model training time is about 30 hours.
  + Each training epoch takes on average 50 minutes on a GPU node with at least 32GB of RAM.
  + Each Morris run with a trained model, and with additional feature analysis that takes around 35 minutes.
* Mean Squared Error (MSE) is used as the loss function.
  + Other metrics include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Symmetric Mean Absolute Percentage Error (SMAPE), and Normalized Nash-Sutcliffe Efficiency (NNSE)
* We clean the outliers from our input features using the following lower and upper thresholds:

A picture containing text, watch, gauge

Description automatically generated

* We compare our TFT model performance with a simple PyTorch Baseline model that predicts future cases and deaths by repeating the last known observations.
* We partitioned our dataset into train, validation, and testing sets
* Our TFT models take the prior 13 days of data as input and use that to predict both cases and deaths for the next 15 days.

**ATTENTION WEIGHTS AND ANALYSIS**

* Analyzing persistent temporal patterns is a key to understanding the time-dependent relationships present in a given dataset.
* We analyze the variable importance by summing up the weights from the Variable Selection Network (VSN) for each variable across the train set

**Feature Sensitivity Analysis**

* We studied the sensitivity of individual features using the Morris method
* Vaccination and disease spread are the most influential features, while other features have much less influence on the output.
* In this study, we experimented with 2 population segmentation methods: 1) Population segmentation by age, and 2) Population segmentation by industry sectors.

**Conclusion and Future Work**

* We find that TFT performs well in learning temporal patterns from the data
* While it learns long range dependencies, it is less effective in highly dynamic forecasting problems with non-stationary sequences, such as extreme or unknown events.
* Combined with the Morris method, this sensitivity study enables us to look into stratifying the modeling based on population subgroups such as industry and age and see how that affects COVID-19 responses to cases of infection and other important features at the community level