EE 380L

Economic Impacts of COVID-19

Clara Cannon, Hassan Hmedi,
Thomas Hunter, Chenkuan Liu,
Aditya Pendyala, Isidoros Tziotis

Overview

1. Defining the problem
2. Exploring the dataset
3. Building the models
4. Analyzing the results

Effects on Health and the Economy

Covid-19 has tremendous consequences both on Health and the Economy.

Effects on Health and the Economy

Covid-19 has tremendous consequences both on Health and the Economy.

How does Covid-19 influence economic indicators in the USA?

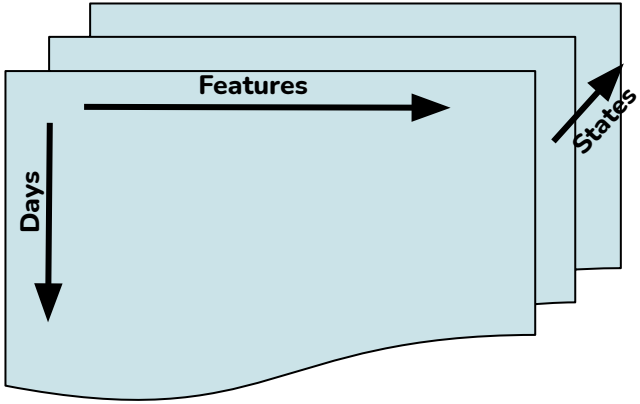
Effects on Health and the Economy

Covid-19 has tremendous consequences both on Health and the Economy.

How does Covid-19 influence economic indicators in the USA?

Can we use covid information to make predictions?

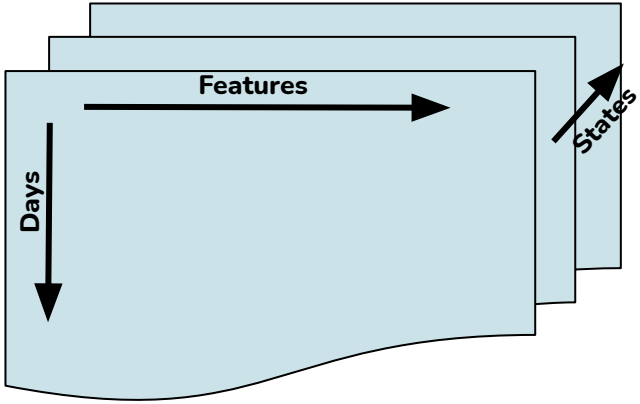
Dataset Overview



Opportunity Insights Economic Tracker: [Chetty et al. 2020]

- Tracks daily employment, spending, revenues, jobs postings, etc. across U.S. counties, industries, and income groups
  - 264 time records x 55 features x 51 states

Dataset Overview

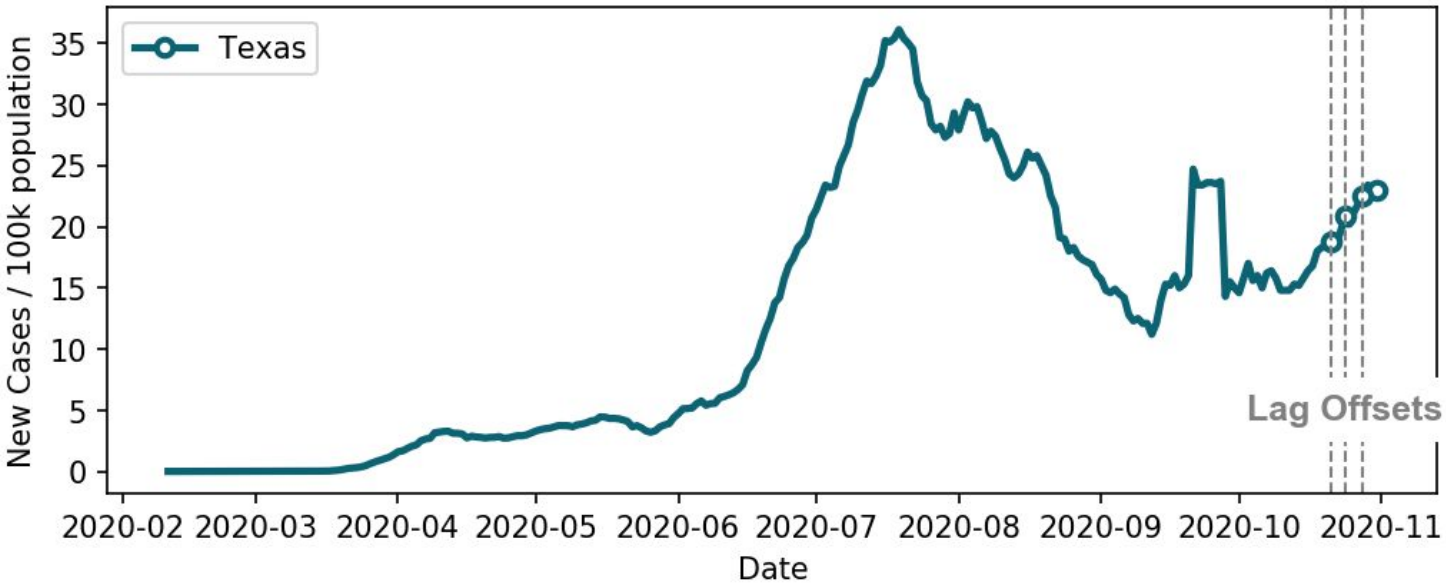


Opportunity Insights Economic Tracker: [Chetty et al. 2020]

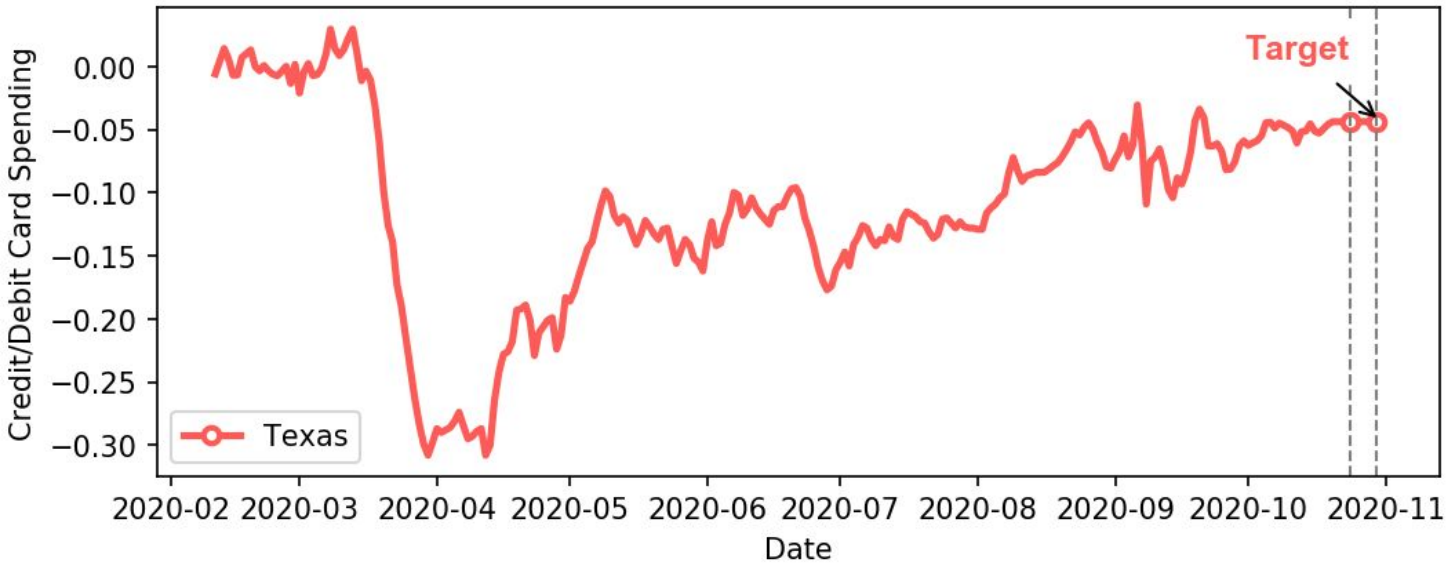
- Tracks daily employment, spending, revenues, jobs postings, etc. across U.S. counties, industries, and income groups
  - 264 time records x 55 features x 51 states

- Compared to data on COVID testing, cases, and deaths
  - 264 time records x 10 features x 51 states

Example Input



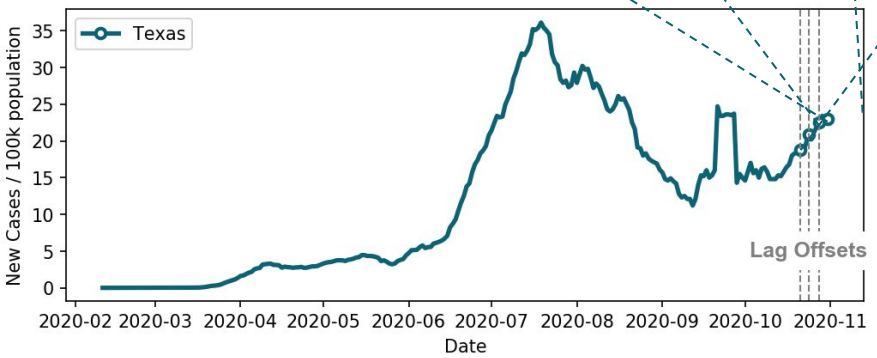
Example Target

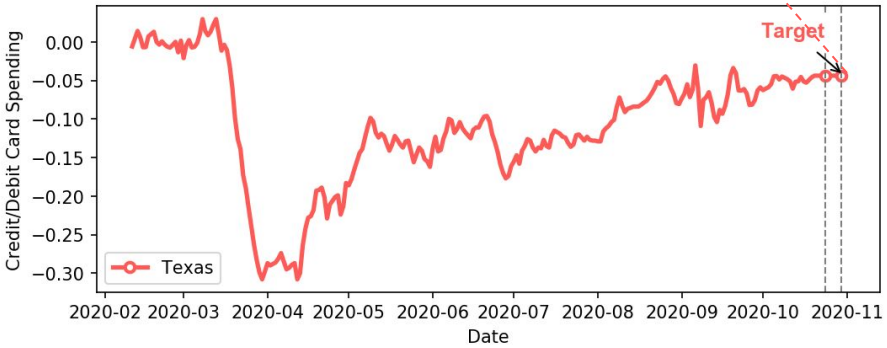


Predictions based on Covid

COVID (Today)	COVID (-3 days)	COVID (-7 days)	COVID (-10 days)	Economy (Today)

Predicting on Covid information.



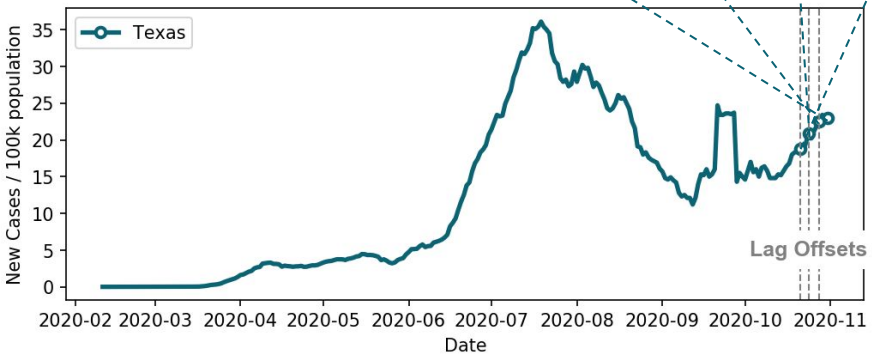


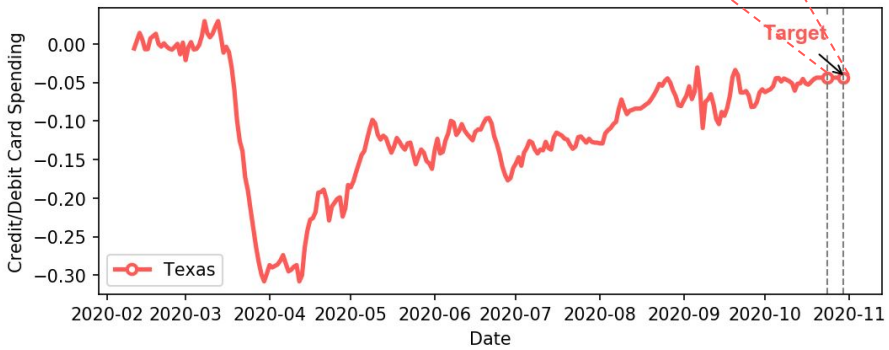
Enhanced Prediction Scheme

COVID (Today)	COVID (-3 days)	COVID (-7 days)	COVID (-10 days)	Economy (Yesterday or -7 days)	Economy (Today)

Predicting on Covid information.

Strengthening the model with past Econ information





Challenges

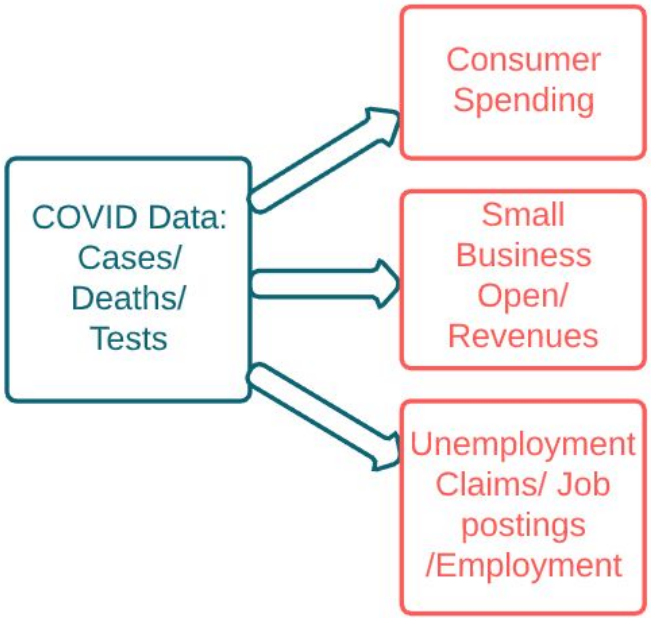
- Time Series Data (Autocorrelated data, Non stationarity)
- Model Choices (Choosing Lag and Training intervals, Expressive subsets of features)
- Missing Values (Which imputation method to use?)
- 51 States (Train a model for each state or Use the state as categorical input?)

Exploring the dataset

Exploratory Data Analysis

- About 70 Features
- Data Categories: Consumer spending, Small Business revenue, Small Businesses open, Employment rate, Covid data, Unemployment Claims, Job Postings
- Data Range: from 01/01/2020 to 11/05/2020
- Data is given as 7 day moving average values in some cases
- Choices Matter!

Flow Chart


COVID Data: Cases/ Deaths/ Tests
Consumer Spending
Small Business Open/ Revenues
Unemployment Claims/ Job postings /Employment

Data Preprocessing

- Data Cleaning
- Imputing
- Target Encoding
- Feature Selection & Engineering
- Time-Series Train-Test Splitting

Data Cleaning, Imputation, & Target Encoding

- Cleaning: Remove January and November
- Imputation: Assign zeros to NAN values in Covid data before March:
  - Absence of Covid cases in most states before March
- Data being time series: time interpolation method has been used
- Target encoding of the states has been used to check if it improves the result

Feature Selection and Engineering

- Select the covid data as input data and the economic data as the target outputs
  a. COVID data: new case rate, new death rate, new positive rate,...
     (Given as absolute values)

  b. Econ data: Spending, Revenues, Employment, Job postings,...
     (Given as fractional change w.r.t their average values in January 2020)

- Hence, all features and outputs are scaled to [0,1]
- Since the data is definitely affected by previous values, delayed versions of both inputs and outputs have been tested.

Input and Output correlation

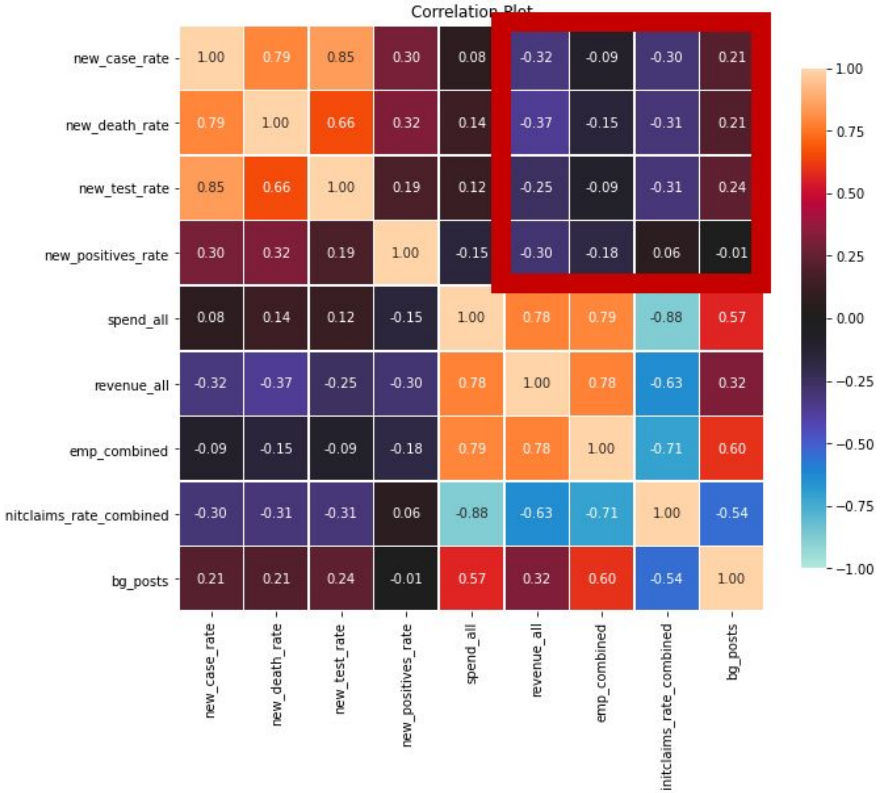


Fig-
Correlation plot
for Texas numbers

Time-Series Train-Test Splitting

- Time Series cross-validator which is a variation of KFold
- Provides train/test indices to split time series data samples that are observed at fixed time intervals, in train/test sets. In each split, test indices must be higher than before, and thus shuffling in cross validator is inappropriate.
- In the k^{th} split, it returns first k folds as train set and the (k+1)th fold as test set.
- Unlike standard cross-validation methods, successive training sets are supersets of those that come before them.

Time-Series Train-Test Splitting, Example

- Assume we have a time series ranging from 2006-2013 and that we set n_splits = 7 then
  - fold 1: training [2006], validation [2007]
  - fold 2: training [2006 2007], validation [2008]
  - fold 3: training [2006 2007 2008], validation [2009]
  - fold 4: training [2006 2007 2008 2009], validation [2010]
  - fold 5: training [2006 2007 2008 2009 2010], validation [2011]
  - fold 6: training [2006 2007 2008 2009 2010 2011], validation [2012]
  - fold 7: training [2006 2007 2008 2009 2010 2011 2012], validation [2013]

Building the models

K-nearest Neighbors Regressor

- Input data: covid data along with delayed versions of it by 3, 7 , and 10 days
- Output: econ data
- Target encoding of the states have been used
- Training set : data between February and July
- Test set: data over the month of August
- Grid Search of the parameters has been used
- Time Series Split: n_splits=5

K-nearest Neighbors Regressor Results

Econ Data	R2
Spending in all merchant categories	0.4971
Net Revenues	0.5755
Small businesses open	0.5961
Employment level for all workers	0.7089
Count of initial unemployment insurance claims	-0.0477

Econ Data	R2
Job postings	-0.4024
Spending on grocery and food stores	0.5537
High income people spending	0.6125
Revenues in leisure and hospitality businesses	0.6721
Education and health services' business open	0.6491

* Preliminary Results

K-nearest Neighbors Regressor Results

Econ Data	R^2
Spending in all merchant categories	0.4971
Net Revenues	0.5755
Employment level for all workers	0.7089

* Preliminary Results

Random Forest Regressor

- Input data: covid data along with delayed versions of it by 3, 7 , and 10 days
    - Target encoding of the states have been used
- Output: Econ data
- Training set : data between February and July
- Test set: data over the month of August
- Grid Search of the parameters has been used
- Time Series Split: n_splits=5

Random Forest Regressor Results

Econ Data	R^2
Spending in all merchant categories	0.7152
Net Revenues	0.7024
Small businesses open	0.7164
Employment level for all workers	0.7276
Count of initial unemployment insurance claims	-0.2348

Econ Data	R^2
Job postings	-0.2695
Low income people spending	0.7144
Employment levels in professional and business services	0.8822
Revenues in leisure and hospitality businesses	0.8506
Transportation business open	0.7509

* Preliminary Results

Random Forest Regressor Results

Econ Data	R^2
Spending in all merchant categories	0.7152
Net Revenues	0.7024
Employment level for all workers	0.7276

* Preliminary Results

Support Vector Regressor

- Input data: covid data along with delayed versions of it by 3, 7 , and 10 days
- Output: econ data
- Target encoding of the states have been used
- Training set : data between February and July
- Test set: data over the month of August
- Grid Search of the parameters has been used
- Time Series Split: n_splits=5

Support Vector Regressor Results

Econ Data	R^2
Spending in all merchant categories	0.113
Net Revenues	0.5004
Employment level for all workers	-0.2819

* Preliminary Results

Support Vector Regressor Results

Econ Data	R2
Spending in all merchant categories	
Net Revenues	
Small businesses open	
Employment level for all workers	
Count of initial unemployment insurance claims	

Econ Data	R2
Job postings	
Low income people spending	
Employment levels in professional and business services	
Revenues in leisure and hospitality businesses	
Transportation business open	

* Preliminary Results

Long Short-Term Memory

[Hochreiter and Schmidhuber, 1997]

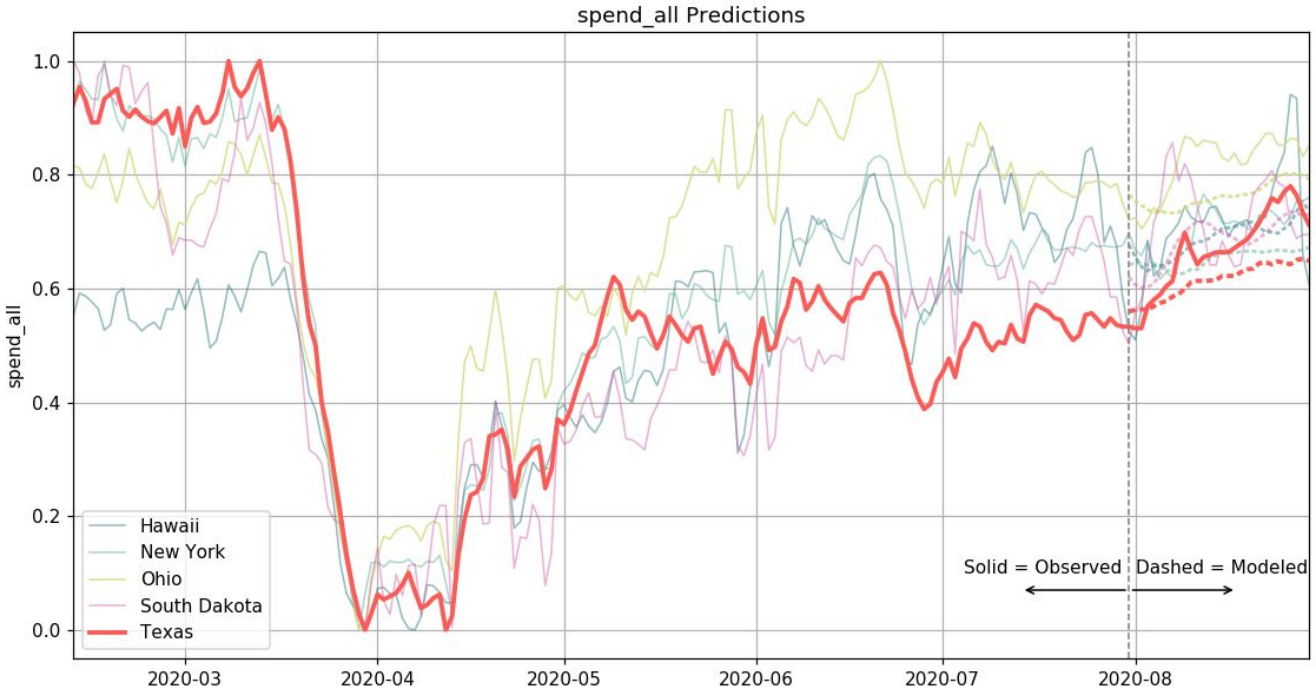
- A modification of RNN which can learn over extended time intervals
- Uses memory cells and “gate units” to filter out irrelevant inputs and outputs during training
- Parameters:
  - Number of hidden layers
  - Size of hidden layers
  - History length

LSTM Results

	COVID Only	COVID + Economy (1 day lag)	COVID + Economy (7 day lag)
Spending	-0.056	0.753	0.436
Revenue	-0.186	0.683	0.465
Employment	0.392	0.947	0.839

* Preliminary Results

Results


spend_all Predictions
spend_all
1.0
0.8
0.6
0.4
0.2
0.0
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
Hawaii
New York
Ohio
South Dakota
Texas
Solid = Observed
Dashed = Modeled

Analyzing the results

Findings

1. COVID intensity not directly correlated with economic trends
    - **Implication:** Policy choices and public sentiment are important to understanding economy
2. Economic data is highly correlated to prior days
    - **Implication:** Near-real time economic health indicators are crucial to understanding the state of economy

Pros/Cons of Models

- KNN Regressor
  - ✓ Simple to interpret
  - ✓ Only one hyperparameter
  - ✗ No ability to deal with missing values
  - ✗ Sensitive to outliers
- RF Regressor
  - ✓ Effective method for estimating missing data
  - ✗ Cannot predict beyond the range in the training data
- LSTM
  - ✓ Works for arbitrary sequence memory
  - ✓ Effective for non-stationary data
  - ✗ Difficult to build and expensive to train

Next Steps

- LSTM development
  - Exploring GPU resource availability
- Apply federated learning techniques
- More vigorous trend and seasonality analysis on dataset
  - Lead to more informed interpolation methods
- Explore other features that might be better at predicting economic trends

References

"The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data", by Raj Chetty, John Friedman, Nathaniel Hendren, Michael Stepner, and the Opportunity Insights Team. November 2020. Available at: https://opportunityinsights.org/wp-content/uploads/2020/05/tracker_paper.pdf

Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.