Photo: MIT Sloan Management Review

Economic Impacts of COVID-19

Chenkuan Liu
Dec 10 · 15 min read

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

Summary

Goal: Understand the connection between COVID-19 spread and the state of the economy

Data: COVID-19 cases, deaths, and testing data; credit card spending, business revenue, and employment

Method: Time-series analysis conducted on state-level data by lagging input and target variables by 1–3 weeks. Models chosen include ensembles, boosted methods, and recurrent networks.

Github link: https://github.com/c3cannon/data_mining_proj

Abstract

Over the past 11 months, the detrimental impact of COVID-19 on lives both in terms of health and economy has been witnessed. In this work, the connections between the trends observed in COVID-19 epidemiology and the state of the economy in the United States are examined. The information provided by the Opportunity Insight Economic Tracker (Chetty et al. 2020) is utilized to train several models based on COVID information from February until July 2020 to derive accurate predictions on specific economic indicators for the month of August. Due to the results, it seems that COVID data can be used to predict these economic indicators with reasonable accuracy.

. . .

Introduction

Since the first COVID-19 (Coronavirus) case confirmed in Wuhan, China, in December 2019, the outbreak continues to spread all across the world. On January 30, 2020, WHO declared the pandemic as an international concern of public health emergency. The novel coronavirus (SARS-CoV-2) disease spread to more than 190 countries infecting more than 66 millions of individuals and causing more than 1,536,000 deaths by December 09, 2020. Although in such times the greatest concern is saving human lives, the next objective is preserving the welfare of the society including the economy. In recent history, it is possible to observe the impact of the Spanish flu (1918–1919) on the economy. Even though the economic data from the early 20th century are scarce, it has been noted that the impact of business closures led to unemployment and the businesses that survived suffered huge losses. Similar comparisons can be made with pandemics from the recent past. During SARS (severe acute respiratory syndrome) in 2003, which lasted less than a year, businesses saw enormous decreases to revenue. A similar

scenario happened in 2009 when the expansion of the H1N1 flu triggered numerous consequences. Likewise, COVID-19 will surely have lasting effects on the global economy and a tremendous impact on the financial markets.

Over the last decade Artificial Intelligence algorithms have been proven successful in solving problems from various fields of science and technology as well as making accurate predictions on the performance of financial markets. Furthermore, recent studies identified that healthcare providers are successfully employing Machine Learning and Artificial Intelligence as they result in better speed, scaling, and reliability (T. Davenport, R. Kalakota, 2019). Therefore, healthcare industries and clinicians worldwide employed various ML and AI methods to tackle the COVID-19 pandemic and address the challenges during the outbreak.

Related work

There has been a concentrated effort in the Machine Learning community to analyze incoming COVID-19 data in order to better understand the hazards and help mitigate the effects of this pandemic. Numerous models have been trained and predictions both on healthcare and the economy have improved the response of our society to the virus. The authors in (Lalmuanawma, Hussain, and Chhakchhuak 2020) present a review of the impact of Machine Learning methods in the arena of screening, predicting, forecasting, contact tracing, and drug development for COVID-19. In the work of (Singh, Kumar, and Sonali 2020) the focus was shifted towards predicting the future reachability of the virus across the United States. Predictions on economic indicators were reported in (Ou et al., n.d.), where the future demand on motor gasoline was predicted while the impact of governmental intervention was also measured. Finally, in (Štifanić et al. 2020) the authors trained a model in order to make accurate predictions on the prices of numerous U.S. stocks.

Contribution

In this work, connections between COVID-19 data and economic indicators such as credit card spending, employment and revenues are examined. The dataset is derived from the Opportunity Insight Economic Tracker (Chetty et al. 2020), where daily records on 10 COVID-19 features are provided along with daily records on 55 economic features. The records cover January up until November of 2020 and are provided for each state in the United States. The main goal is to utilize machine learning techniques

in order to train efficient models based on COVID information that make reliable and accurate predictions about the future of the economy across the states. Special techniques are utilized to carefully manipulate time series while various results are presented based on different information and optimization schemes. Standard baseline methods were compared such as Decision Trees, Nearest Neighbors, Random Forests as well as more advanced methods such as XGboost and Recurrent Neural Networks. Finally, findings are analyzed and insights on the behavior of models based on COVID information and governmental intervention are provided.

. . .

Dataset Description

The dataset is acquired from the Opportunity Insights Economic Tracker database which aggregates anonymized data from credit card and payroll firms along with statistics on the spread of COVID-19.

Target Variables

The target variables considered include the following:

Spending, Small business Revenue, Small businesses open, Employment rate, Unemployment Insurance claims, Job postings.

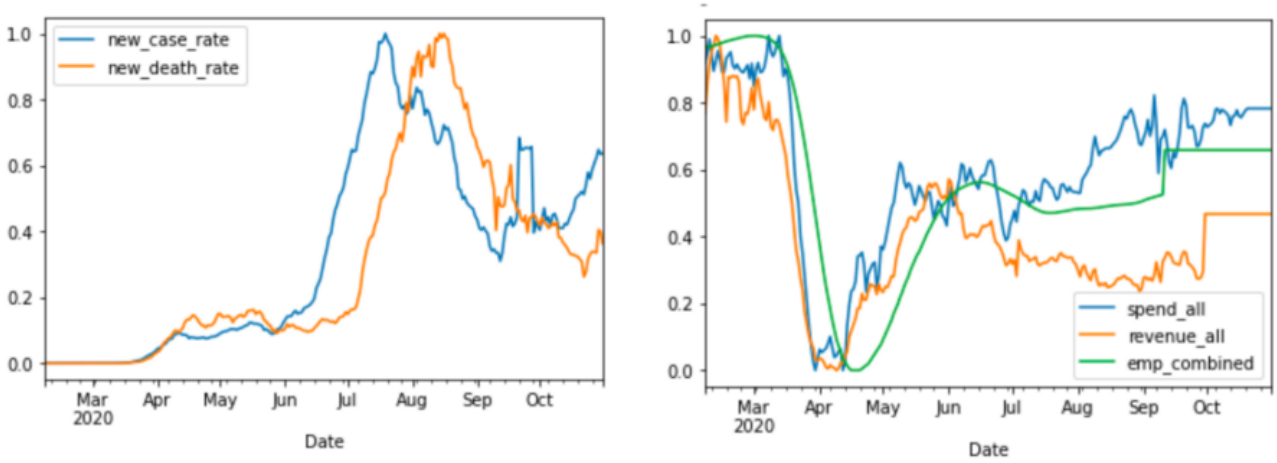
The dataset also includes subcategories of all these indicators, which are broadly based on the economic sectors or the income level of the individual (e.g. spending on entertainment/merchandise/education, spending by low/high income individuals, employment for workers in transportation sector, revenue earned by small businesses in health/education sectors).

The economic data barring unemployment claims is presented as percent change in the respective values as compared to the month of January, when there wasn't any COVID effect as such. Unemployment claims are given as the number of claims per 100 people. Most of the target variables are provided as 7-day moving averages to smooth out the spikes and account for weekly patters.

Input Variables

The input features are constituted by the COVID information which includes rates of cases, deaths, tests, positive tests at daily and cumulative scale. Rates are values per 100k population. These too are given as 7-day moving averages.

Exploration



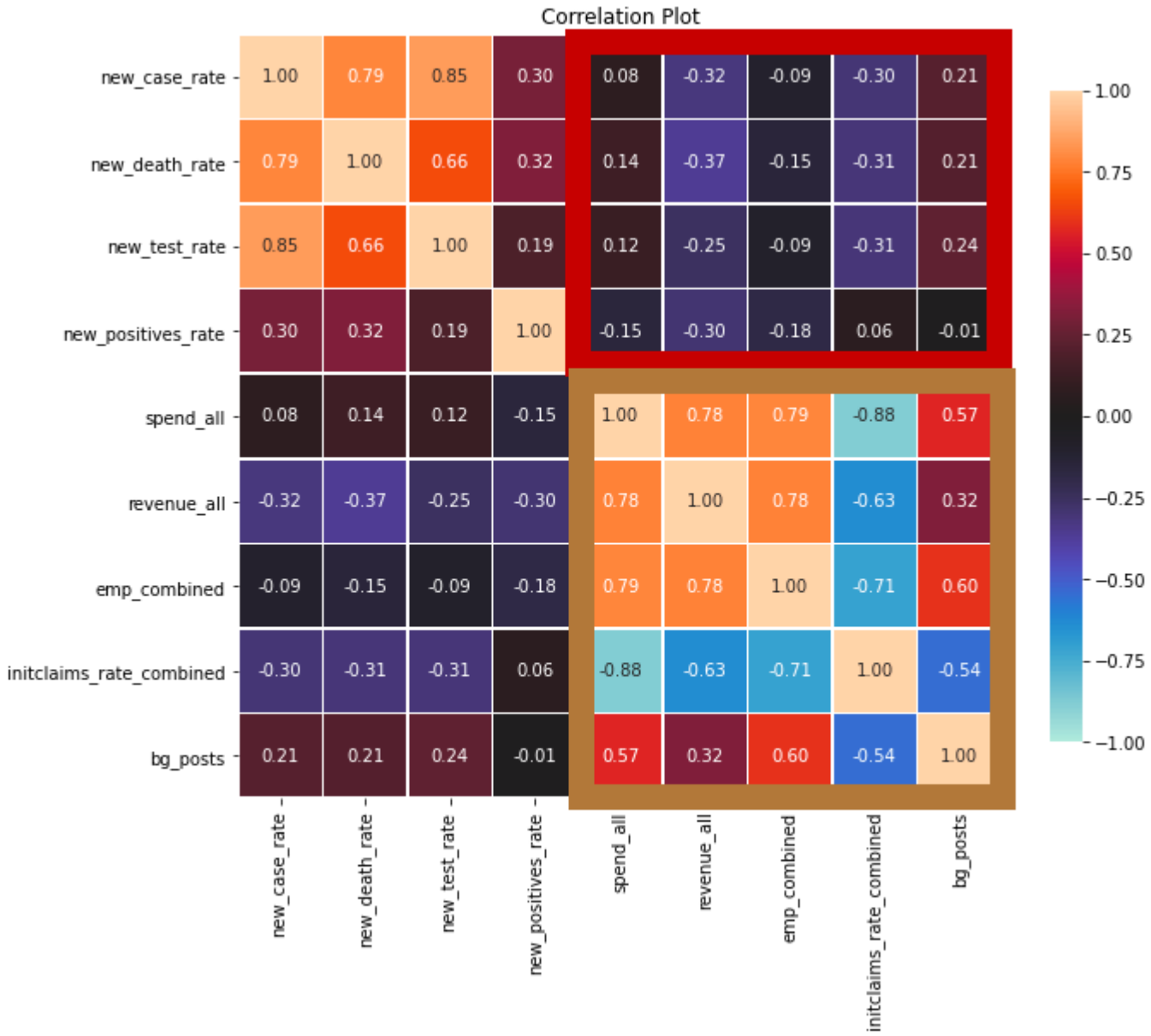
Sample data for 2 input and 3 target variables for the state of Texas

Many data values were missing in the dataset. To fill those in, the time interpolation method had been used `df_simple = df_simple.interpolate(method = 'time')`. Also, the values in January were all 0 for the COVID data. These records would be outliers in the context of our problem as we are only interested in the trajectory once COVID started. Therefore, only records starting from February are considered.

While deciding to do this project, we hypothesized a strong correlation between past values (say, a week or two) of COVID and the current values of these economic indicators. As it can be seen from the time series plots above, this is clearly not the case. This reduces the hope of getting good performance from baseline models like Linear Regression. The economic indicators nosedived in early April due to panic and lockdowns and then slowly started recovering, even though cases were continually increasing.

However, significant correlation exists between the set of target variables. Consider the following correlation plot among the various time-series. The top right rectangle contains correlation between inputs and outputs. Their magnitudes are pretty close to 0.

On the other hand, the correlation among the economic variables is pretty strong (bottom right), as expected.



Pearson Correlation Matrix (Data from state of Texas fips = 48)

```
pearson_corr = dataframe_collection[48]
[['new_case_rate_14','new_death_rate_14','new_test_rate_14','new_posi
tives_rate_14','spend_all','revenue_all','emp_combined','initclaims_r
ate_combined','bg_posts']].corr(method='pearson')

fig, ax = plt.subplots(figsize=(10,10))
plt.title('Correlation Plot')

sns.heatmap(pearson_corr, vmin=-1.0,vmax=1.0, center=0, fmt='.2f',
square=True, linewidths=.5, annot=True, cbar_kws={"shrink": .70})
```

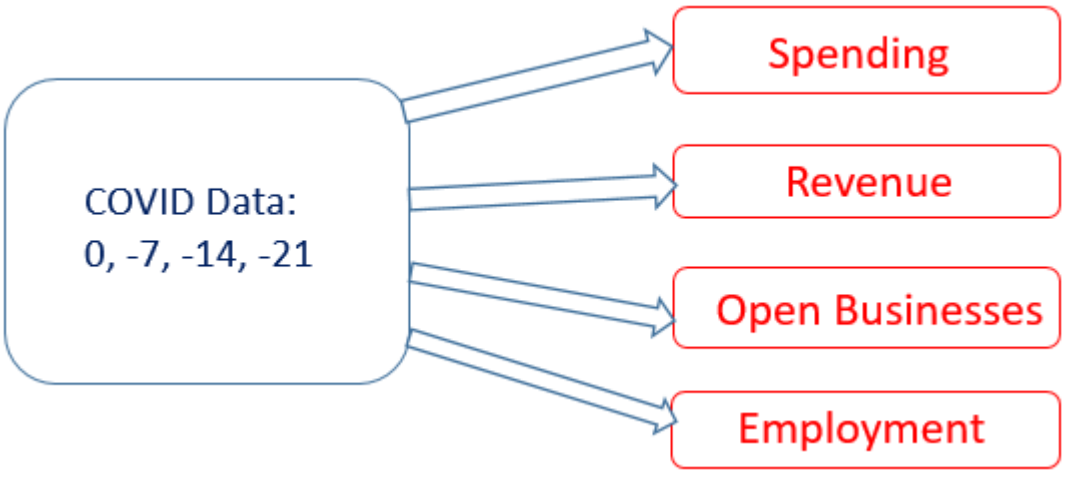
```
plt.show();
```

Feature Engineering

As far as feature selection is concerned, the cumulative values of COVID numbers are redundant. From the theory of adjusted R^2, these numbers will only deteriorate the performance on the test set. Hence, only daily values of COVID information were considered.

The data at hand is for the 51 jurisdictions representing the U.S. states and the District of Columbia. There are several options for addressing this variable: train a separate model for each state, drop the state label and seek a common model that relies only on past data, or convert the state feature into an input feature. The latter option was chosen by employing target encoding which replaces a categorical feature with the mean of the target. For example, when predicting total_spending, simply replace the state label with the total_spending's mean over that state in the training set.

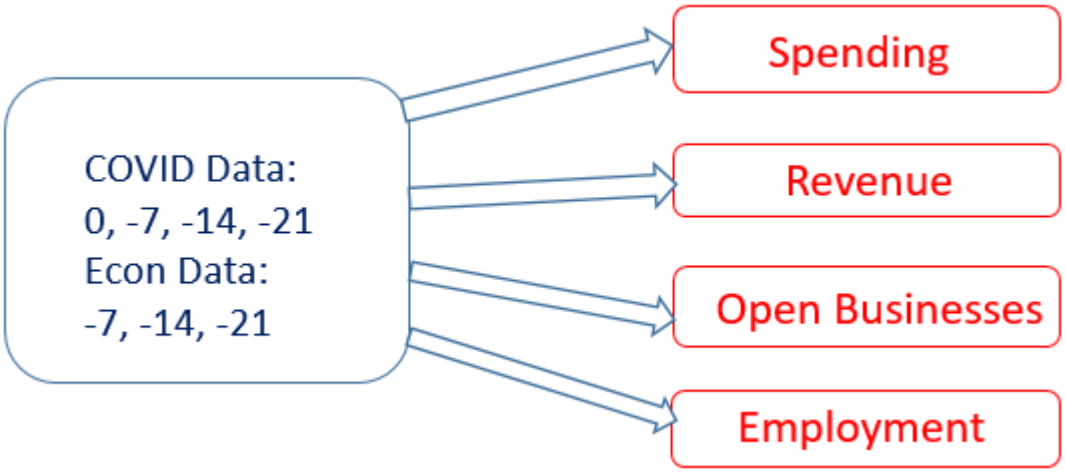
Since the input set consists of only 11 features, feature reduction methods like PCA are not recommended. Scaling, which under certain models improves performance and convergence rate, was employed through `StandardScaler()`. As stated before, we wanted to incorporate the idea of causality, so shifted values of COVID was used for prediction as shown below. These shifts were varied to get better results.



Predicting Econ Data using current COVID Data and shifted versions of it by 1,2,3 weeks

Later, we considered a related problem where the past economic (target) data is used. This led to an enhanced prediction scheme that may more accurately reflect data that

would be available to businesses and policymakers hoping to predict and mitigate economic challenges.



Predicting Econ Data using current COVID Data and shifted versions of COVID and Econ by 1,2,3 weeks

. . .

Learning/Modeling

Selection of Models

To study the problem, we began with simple models and added complexity until we were satisfied with the results. Our initial test was a Multivariate Linear Regression (MLR) model, which returned a negative R^2 correlation coefficient. Following this, we tried Neural Network and Support Vector Regression models, which also returned negative R^2 values. Our hypothesis at this point was that the high variance in the target data along with the high correlation of input variables was limiting the efficacy of standard models. With this in mind, we choose to build ensemble, boosted, and recurrent models. Models with promising initial results were then further improved by the selection of hyperparameters, while models with poor initial results such as MLR were dropped. Our final selection of models includes a Long Short-Term Memory (LSTM) network, Random Forest, Nearest Neighbors, Catboost, etc.

Training Methods

Since the data at hand is a time series data, K-Fold cross validation cannot be used. Instead, we used the function `TimeSeriesSplit` from scikit-learn since keeping track of

the time indices is crucial to our experiments. When it comes to hyperparameter selection the following models along with their respective hyperparameters have been used:

```
models = [RandomForestRegressor(),KNeighborsRegressor(),
xgb.XGBRegressor(objective='reg:squarederror', random_state=42),
LGBMRegressor(), CatBoostRegressor(verbose=False)]

param_RF = {'n_estimators': [20],'max_depth' : [5,10,15] }
param_KNR = { 'n_neighbors':[5,10,15,20,25] }
param_XGB = {}
param_Cat = {}
param_lgbm = {}
params = [param_RF, param_KNR, param_XGB, param_lgbm, param_Cat]

gsearch = GridSearchCV(estimator=models[i], cv=tscv,
param_grid=params[i], scoring = 'r2')

gsearch.fit(X_train, y_train)
best_score = gsearch.best_score_
best_model = gsearch.best_estimator_
```

Design Choices:

As discussed above, we used a target encoding of the states. For each target variable, we applied the model separately using the encoded target mean in place of the state.

```
for column in df_econ_new.columns:
  if column != 'statefips':
    means = df_econ_new.loc[:'2020-07'].groupby('statefips')
[column].mean()
    df_econ_encoded = df_econ_new.copy()
    temp = df_econ_encoded['statefips'].map(means)
    df_covid_encoded = df_covid_new.copy()
    df_covid_encoded['statefips'] = temp
```

In addition, since the variations in the economic data is also affected by the COVID data at previous time stamps, a shift in the input data must also be used. The function used for shifting the data is the following

```
def df_derived_by_shift(df,lag=0):
   df = df.copy()
   if not lag:
       return df
   cols ={}
   for i in lag:
       for x in list(df.columns):
              if not x in cols:
                     cols[x] = [‘{}_{}’.format(x, i)]
              else:
                     cols[x].append(‘{}_{}’.format(x, i))
   for k,v in cols.items():
      columns = v
      dfn = pd.DataFrame(data=None, columns=columns, index=df.index)
      i = 0
      for c in columns:
         dfn[c] = df[k].shift(periods=lag[i]*51) ## this is
because we want to shift all the 51 states\
         i = i+1
      df = pd.concat([df, dfn], axis=1)
return df
```

We proposed two design choices where we vary the input space:

1. Use COVID data along with shifted versions of it by 7, 14, and 21 days
2. Use the same data in step (a) along with shifted versions of the economic data by 7, 14, and 21 days

```
delay_covid = [7,14,21]
delay_econ = [7,14,21]

df_covid = df_derived_by_shift(df_covid, delay_covid)
df_econ = df_derived_by_shift(df_econ, delay_econ)
```

The training set was selected to be the data between the months of February and July and the test set was the month of August. Since we have 6 months of training data, we used `tscv = TimeSeriesSplit(n_splits=5)` since the training sets in this case would be better interpreted.

. . .

Results

Model Evaluation

In the following table we summarize the main models results we obtained in this project.

R^2					
	Model Name	**Spending**	**Revenue**	**Merchant**	**Employment**
Results using COVID data Only	LSTM	0.4319	0.2590	0.3356	0.3628
	RF	0.7409	0.6400	0.7152	0.7041
	KNR	0.4953	0.6168	0.5991	0.7142
	XGBoost	0.7378	0.6580	0.6757	0.7063
	LGBM	0.7685	0.7171	0.7463	0.7251
	CatBoost	0.7761	0.7314	0.7510	0.7714

R^2					
	Model Name	**Spending**	**Revenue**	**Merchant**	**Employment**
Results using COVID and past Econ data	LSTM	0.7754	0.8341	0.8440	0.9289
	RF	0.7235	0.8202	0.8268	0.9415
	KNR	0.7508	0.7227	0.6958	0.8550
	XGBoost	0.7759	0.8320	0.8504	0.9768
	LGBM	0.7985	0.8274	0.8524	0.9654
	CatBoost	0.8064	0.8449	0.8508	0.9760

Summary of model performance. Adding Econometric Data increases R^2 in most cases.

Several metrics have been used to evaluate the performance of the chosen models through the following function:

```
def regression_results(y_true, y_pred):
  #Regression metrics
  explained_variance=metrics.explained_variance_score(y_true,
y_pred)
  mean_absolute_error=metrics.mean_absolute_error(y_true, y_pred)
  mse=metrics.mean_squared_error(y_true, y_pred)
  median_absolute_error=metrics.median_absolute_error(y_true, y_pred)
  r2=metrics.r2_score(y_true, y_pred)

  print('explained_variance: ', round(explained_variance,4))
  print('r2: ', round(r2,4))
  print('MAE: ', round(mean_absolute_error,4))
```

```
    print('MSE: ', round(mse,4))
    print('RMSE: ', round(np.sqrt(mse),4)
```

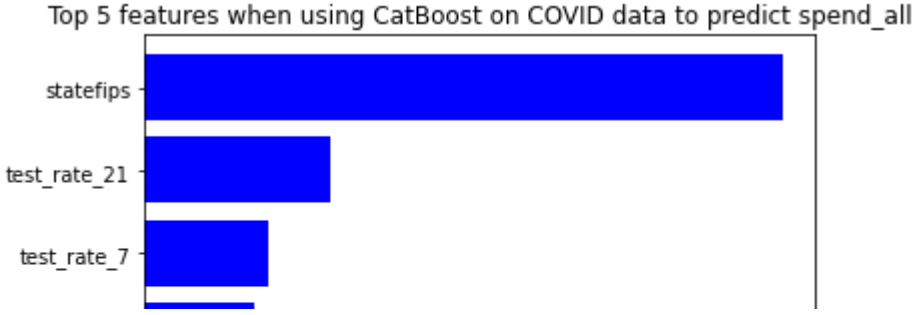
Model evaluation was done using the R^2 metric. Other metrics such as mean square error, mean absolute error, and root mean square error were evaluated and the results can be found in the `.txt` files on Github. The reason that R^2 is the significant metric is the fact that the target output values are small in magnitude. The magnitude of the target variables of interest is of order 10^{-2} so a MAE of 0.03 for example is very large and cannot be used.

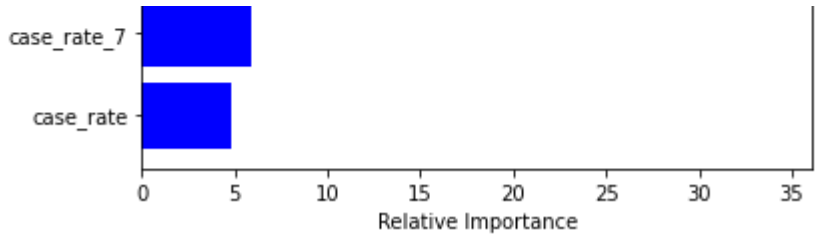
In the table above, it is notable that while the LSTM model performs well when economic data is included in the input, it struggles in comparison to other models when it only has COVID data. This type of model is especially suited for time-series analysis where the past values of the target variable are known.

Using the R^2 metric, we can see that CatBoost is the model which outperformed the other models presented in the table above. As we can also see, adding shifted versions of the Econ Data improves the results significantly in all the cases except when using Random Forest to predict the total spending.

Insights

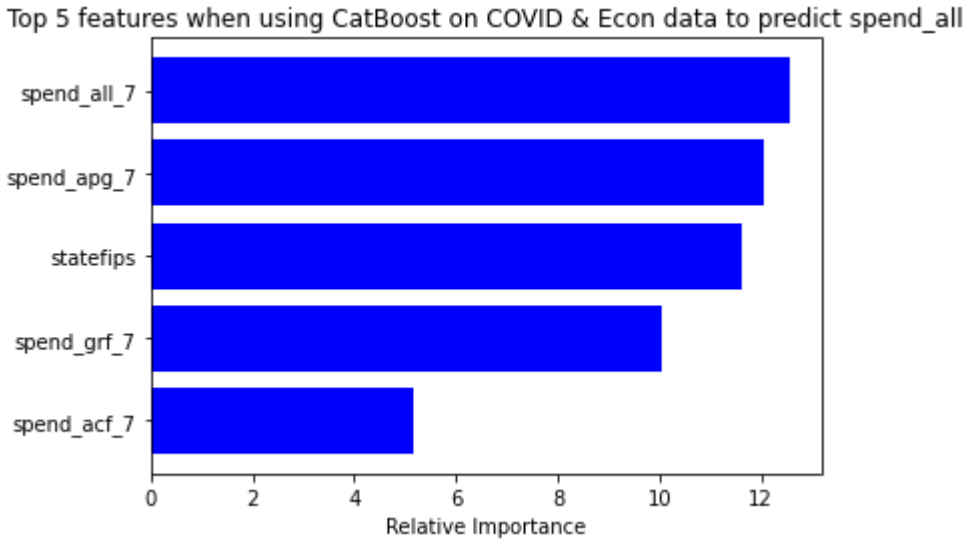
We ultimately decided on using target encoding of the state variable. We explored a model which was agnostic of the state. To standardize each state, the input and target variables were scaled to the minimum and maximum values for each state separately. The model was then trained with only knowing the scaled input variables and corresponding lagged variables, but had no knowledge of which state the data point corresponded to. We found that this model performed significantly worse. This finding suggests that something besides just COVID spread, such as state policies or public behaviors towards the virus, have an effect on the economy.





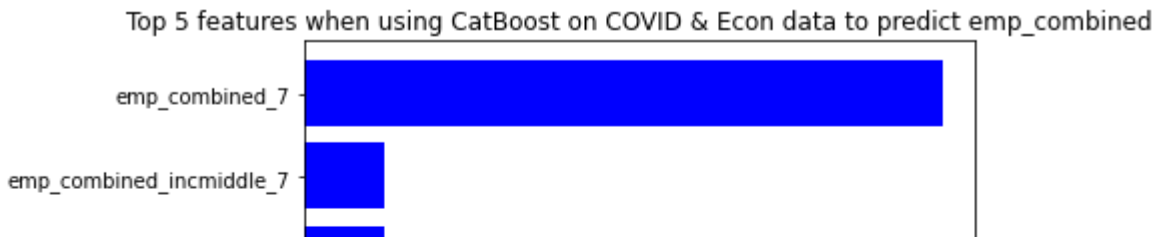
When only COVID data is input, the most important feature is the state encoded variable

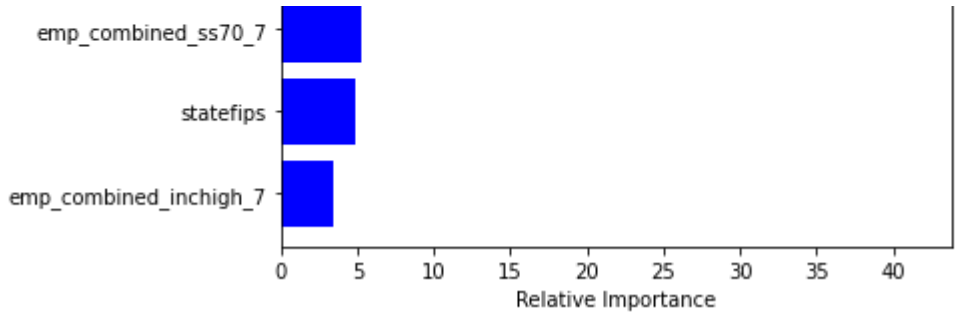
To investigate the data further, we examine the importance of input features on the model outcomes. Using Catboost as an example, we can see that the state classification plays a very large role in the predictor when only the COVID data is included. As the `statefips` has been encoded, its value contains the mean of the `spend_all` variable for that state. Therefore, the high importance assigned implies the model has a high bias towards the mean of the time-series.



When economic data is included, the previous week's economic activity is very important to the model

Conversely, when econometric data is also included as an input, the most important variables become the lag of the target variable and its subcategories. The model exhibits higher autocorrelation in the time-series.





Stable targets such as employment place more importance on the prior week's target value

We also observe that the more stationary and low variance the target, the greater the significance the model assigns to the prior week's value. In the employment variable, the value remains much more steady from week to week, and therefore the model can primarily rely on prior week's value regardless of state or other factors.

The following is the code used to obtain the top five most important features selected by the tested models:

```
features = list(df_covid_encoded.columns.values)

importances_full = best_model.feature_importances_
indices_full = np.argsort(importances_full)
indices_full = indices_full[len(indices_full)-5:len(indices_full)]

title = 'Top 5 features when using '+ model_names[i]+ ' on COVID &
Econ data to predict '+ str(column)

plt.title(title)
plt.barh(range(len(indices_full)), importances_full[indices_full],
color='b', align='center')
plt.yticks(range(len(indices_full)), [features[i] for i in
indices_full])
plt.xlabel('Relative Importance')
plt.show()
```

. . .

Conclusion

In this project, we explored the connections between COVID-19 trend data and economic indicators in the United States. Not only did our results prove the predictive

power of pandemic data on other time-series domains, we also established a baseline for a relatively new problem in the data mining and analytics field. With limited computational resources and standard statistical data processing techniques, we achieved close to 0.85 R^2 with our strongest model.

Even though our initial exploration showed that COVID-19 data metrics were not strongly correlated with economic features due to the former's high variability and the latter's contrasting stability, we were able to uncover the hidden connections with our selection of trained models. Through this process, we learned the tenuous nature of time-series prediction. Often, we had to look outside of the data set to truly understand the factors influencing the rise and fall of our target variables. The importance of domain knowledge and understanding external circumstances were just as valuable in our decision making as technical prowess. It is easy to go overboard with feature engineering and force data to portray an inaccurate version of reality. All of our efforts focused on letting the data tell its own story.

Future Work

In the future, possible extensions of the project include pre-training more data hungry models, such as the LSTM, on past pandemic data. This could unveil similarities between historic health crises and predict targets with higher confidence. Combining the time series features from this data set with other hand selected features such as temperature, calendar season, and regional policy would also likely boost model performance. Another interesting extension includes adding more interpretability to our results. Explainable artificial intelligence (AI) is a broadening field. It would be intriguing to develop a model capable of explaining its "thinking process" in layman's terms for the general public.

We can also extend the use-case of the model by feeding in hypothetical input data, such as COVID-19 projections from the Institute for Health Metrics and Evaluation (IHME) at the University of Washington. By extending the model in this way, we could allow policy makers to better anticipate the consequences of the pandemic and craft measures to alleviate the negative effects. Likewise, considering the subcategories of economic variables in our dataset would provide insight into industries and population segments which have been impacted most by the pandemic.

• • •

References

Brotherhood, L., Kircher, P., Santos, C., & Tertilt, M. (2020, May). An Economic Model of the COVID-19 Epidemic: The Importance of Testing and Age-Specific Policies. *IZA Institute of Labor Economics*, *13265*. http://ftp.iza.org/dp13265.pdf

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

COVID Economics. (2020). Center for Economic Policy Research. https://cepr.org/content/covid-economics-vetted-and-real-time-papers-0

Darkhorse Analytics. (2020). *Opportunity Insights*. Economic Tracker. Retrieved 10 19, 2020, from https://www.tracktherecovery.org/

Haddadpour, F., Kamani, M. M., Mokhtari, A., & Mahdavi, M. (2019). Federated learning with compression: Unified analysis and sharp guarantees.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). *Artificial Intelligence and Statistics*. PMLR.

Ross, C. P., & Ross, S. Y. (2020, May 26). *Forecasting the Economy During COVID-19*. Yale School of Management. https://som.yale.edu/blog/forecasting-the-economy-during-covid-19

Wang, W., Liu, Q., Liang, H., Joshi, G., & Poor,, V. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization. *arXiv preprint arXiv:2007.07481*.

Chetty, Raj, John Friedman, Nathaniel Hendren, and Michael Stepner. 2020. "The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data." *Opportunity Insights*, (December).

Lalmuanawma, Samuel, Jamal Hussain, and Lalrinfela Chhakchhuak. 2020. "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review." *Chaos, Solitons & Fractals* 139:110059. https://doi.org/10.1016/j.chaos.2020.110059.

Ou, Shiqi, Xin He, Weiqi Ji, Wei Chen, Lang Sui, Yu Gan, Zifeng Lu, et al. n.d. "Machine learning model to project the impact of COVID-19 on US motor gasoline demand." *Nature Energy* 5 (9). 10.1038/s41560–020–0662–1.

Singh, Punn N., Sonbhadra S. Kumar, and Agarwal Sonali. 2020. "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms." *medRxiv*. 10.1101/2020.04.08.20057679.

Štifanić, Daniel, Jelena Musulin, Adrijana Miočević, Sandi B. Šegota, Roman Šubić, and Zlatan Car. 2020. "Impact of COVID-19 on Forecasting Stock Prices: An Integration of Stationary Wavelet Transform and Bidirectional Long Short-Term Memory." *Complexity* 2020:12. 10.1155/2020/1846926.

T. Davenport, R. Kalakota, "The potential for artificial intelligence in healthcare", Fut Healthc J, 6 (2) (2019), pp. 94–98, 10.7861/futurehosp.6–2–94

Economics Data Science Covid 19 Towards Data Science Machine Learning