Can Google Trends Predict U.S. Initial Claims?

The College of William & Mary
Jacob Somer
jsomer@email.wm.edu

I. Introduction

The unemployment rate has been one of the primary indicators for the health of the economy. Used by both policy makers and hedge fund managers alike, it is a key driver for important decisions in the economy. For example, the spike in the unemployment rate during March of 2020 is in part what led the Federal Reserve to slash interest rates to near zero.[1] Therefore, the motivation for this paper is to provide meaningful insights in the art of forecasting unemployment.

As we enter the world of big data and high performance computing, the ability to forecast complex relationships has never been more exciting. With the growing popularity of the internet, there is more data today than there has ever been. This is supported by the fact that there are now 134 phone subscriptions per 100 people in the U.S. alone.[2] In unison with more people depending on the web for their daily routines, big tech companies like Google have reigned supreme with roughly 92% market share in global search queries.[3] Luckily, there exists an open-source tool known as Google Trends. Through this, researchers are able to gain access to high resolution, real-time data on human behavior in cyberspace. This paper aims to test the predictive power of Google Trends search data on weekly Initial Claims in the United States. Leveraging the power of newly available data, and state of the art modeling techniques, I find a positive relationship between google trend searches and the ability to forecast weekly unemployment claims.

I reach this conclusion by comparing the forecasting performance of three models. First, I use a simple persistence model to establish a baseline for prediction. Second, with this baseline, I fit a classic Vector Autoregressive (VAR) model to the dataset. Finally, I compare these two models with the newer Long-Short Term Memory (LSTM) network. After fitting these models, I compare their out-of-sample forecasting accuracy.

II. Previous studies

Prior to this paper, there have been many before me using Google Trends to forecast the unemployment rate. First and foremost, I must credit the incredible work done by Choi and Varian in *Predicting the Present with Google Trends* (2009) for inspiring me to investigate this topic.[4] While working at Google, they demonstrated promising results in forecasting many

---

[1] "Implementation Note Issued March 15, 2020," Board of Governors of the Federal Reserve System, 1-2, accessed May 16, 2021, https://www.federalreserve.gov/newsevents/pressreleases/monetary20200315a1.htm)
[2] "Mobile Cellular Subscriptions (per 100 People)," Data, 1, accessed May 16, 2021, https://data.worldbank.org/indicator/IT.CEL.SETS.P2)
[3] Sourav Kundu and Rajshekhar Singhania, "Forecasting the United States Unemployment Rate by Using Recurrent Neural Networks with Google Trends Data," *International Journal of Trade, Economics and Finance* 11, no. 6 (December 6, 2020) :6, doi:10.18178/ijtef.2020.11.6.679)
[4] Hal R. Varian and Hyunyoung Choi, "Predicting the Present with Google Trends," *SSRN Electronic Journal*, April 10, 2009, 2, doi:10.2139/ssrn.1659302)

different types of time series using a simple AR model and Google Trends data. Later, in Choi and Varian (2012), they apply that same methodology to the U.S. Unemployment Rate. They find that Google Trends did not improve forecasting MAE overall, but it did improve forecasting for the 2009 recession by 13.6 per cent.[5] Other researchers like Askitas and Zimmerman (2009) found evidence that Google Trends improves forecasting accuracy in Germany using an Error Correction Model (ECM) suggesting the possibility of cointegration between keywords.[6] Interestingly, Baker and Fradkin (2011) examined job search frequency in response to unemployment payments. Using a fixed effects model, they found that, “Those with fewer than ten weeks remaining [in recieving unemployment checks] search 66% more than those with ten to twenty weeks remaining and 108% more than those with more than thirty weeks remaining”.[7]

More recently, Maryem Rhanoui et al. (2019) compared the well-known ARIMA model to a Long-Short Term Memory network on various univariate time series (stock-market data, exchange rates, housing prices, etc...). According to the authors, “the average reduction in error rates obtained by LSTM is between 84 - 87 percent when compared to ARIMA indicating the superiority of LSTM to ARIMA”.[8] More specific to the data used in this study, Werken and Smit (2019) found evidence that recurrent neural networks perform well in forecasting inflation and unemployment in the univariate case.[9] Kundu and Singhania (2020) also showed promising results with neural networks. Specifically, they found that the LSTM performed better than a baseline VAR when forecasting monthly unemployment with Google Trends.[10] Much of this paper aims to reinforce the findings of Choi and Varian relating google trends to unemployment claims while experimenting with a new model architecture: the LSTM.

III. Theory

Every Thursday at 8:30 a.m., the United States Department of Labor releases a report that includes the number of people who filed for unemployment insurance that week.[11] It has been noted by various economists that weekly unemployment claims have a strong relationship with cyclical troughs in the U.S. economy. Historically, unemployment claims peak several months before the bottom of a recession.[12]

---

[5] Hyunyoung Choi and Hal Varian, "Predicting the Present with Google Trends," *Economic Record* 88 (June 2, 2012): 5, doi:10.1111/j.1475-4932.2012.00809.x)
[6] Nikos Askitas and Klaus F. Zimmermann, "Google Econometrics and Unemployment Forecasting," *SSRN Electronic Journal*, April 2009, 2, doi:10.2139/ssrn.1480251)
[7] Scott R. Baker and Andrey Fradkin, "What Drives Job Search? Evidence from Google Search Data," *SSRN Electronic Journal*, 2011, 3, doi:10.2139/ssrn.1811247)
[8] Maryem Rhanoui et al., "Forecasting Financial Budget Time Series: ARIMA Random Walk vs LSTM Neural Network," *IAES International Journal of Artificial Intelligence (IJ-AI)* 8, no. 4 (2019): 2, doi:10.11591/ijai.v8.i4.pp317-327)
[9] Werken, Laurens & Smit, Victoria. (2019). Exploring the Use of Recurrent Neural Networks for predicting Inflation and Unemployment. 10.13140/RG.2.2.29219.81446.
[10] Sourav Kundu and Rajshekhar Singhania, "Forecasting the United States Unemployment Rate by Using Recurrent Neural Networks with Google Trends Data," *International Journal of Trade, Economics and Finance* 11, no. 6 (December 6, 2020): 1, doi:10.18178/ijtef.2020.11.6.679)
[11] https://www.dol.gov/ui/data.pdf
[12] https://research.stlouisfed.org/publications/economic-synopses/2018/06/01/recession-signals-the-yield-curve-vs-unemployment-rate-troughs/

When someone becomes unemployed, it is expected that they search for things related to welfare and jobs. For example, during the spike in unemployment claims in March 2020, Florida's state unemployment benefits website crashed.[13] This implies that internet searches for Florida's Department of Economic Opportunity must have increased above normal that month. For this reason, it seems likely that welfare related Google trends would rise in correlation with unemployment claims. I verify this in figure 2 with the correlation matrix between certain keyword searches and unemployment claims.

Another thing to note is that Google Trends data, in theory, can be accessed instantaneously. That is, if it is Monday, we can create a nowcast of unemployment claims before the data is officially released on Thursday. Again, this assumes we have a good model for making the prediction.

IV. Data

One of the greatest challenges in modeling with Google trends is choosing the right keywords. Luckily, Choi and Varian (2012) provide a substantial list that is easy to interpret. In their paper they extracted keywords from two main categories: *Welfare & Unemployment* and *Job Search*. From *Welfare & Unemployment*, the keywords were *Social Security, Social Security Office Locations, Unemployment Benefits, Social Security Gov, Unemployment Office, Food Stamps, Department of Labor*. From the *Job Search* category, the keywords were *Monster, Indeed, Jobs, Job Search, Resume, Job Search Engines, Linkedin, Hotjobs, Cover Letter*.[14] I used a python API to extract Google Trends weekly interest for each keyword. Figure 1 depicts an example of a few keywords from the dataset along with U.S. initial claims. The full dataset can be found here: *https://github.com/jacobsomer/Google-Trends-and-Unemployment/tree/main/data*.

Creating a separate coefficient for each keyword would lead to a model that is hard to interpret (especially when we add lags). To solve this, I simplified the dataset by taking the average of each keyword in the two main categories. That way, we can create a model similar to Choi and Varian. The final dataset has 783 observations of three variables: U.S. Initial claims, *Welfare & Unemployment*, and *Job Search*. Based on figure 1, it is clear that during the month of March 2020, there was an anomalous spike in both the trends and the unemployment claims level. We can also see from the correlation matrix in figure 2, that there is a strong correlation between unemployment claims and the Google search for *Unemployment Benefits*. As there are no major warning signs at first glance, we proceed with modeling the data to achieve a more precise understanding.

[13] https://www.wtvy.com/content/news/Florida-unemployment-website-crashes-again-new-unemployment-data-county-by-county-570005621.html

[14] Hyunyoung Choi and Hal Varian, "Predicting the Present with Google Trends," *Economic Record* 88 (June 2, 2012): 5, doi:10.1111/j.1475-4932.2012.00809.x)

Figure 1.

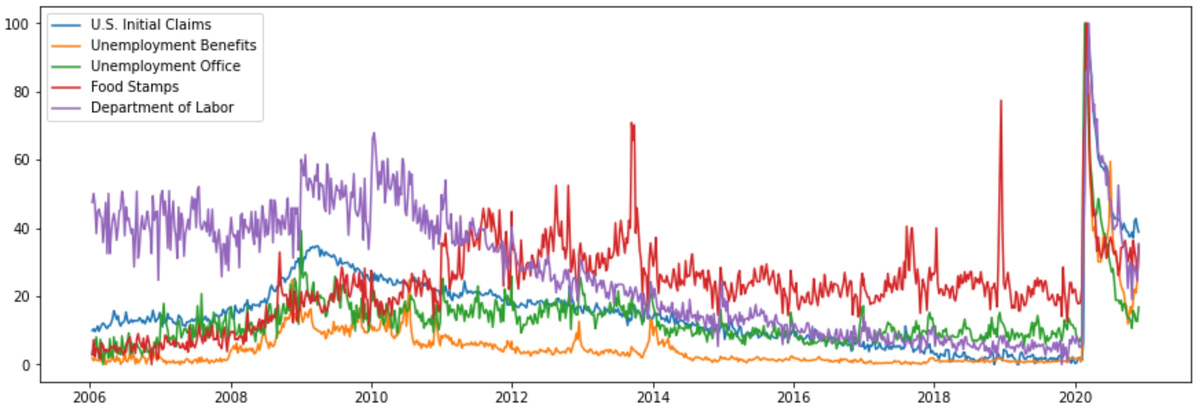
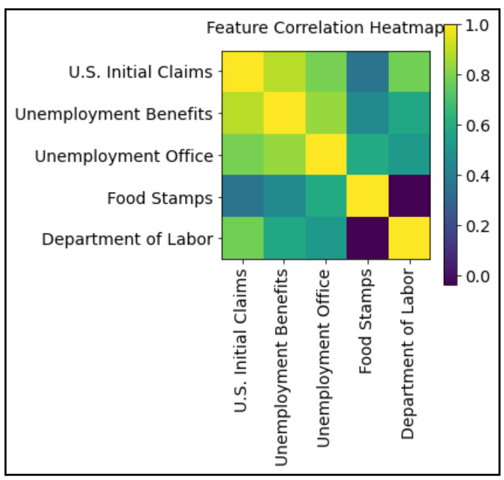


Figure 2.



V. Model

A. Vector Autoregressive Model (VAR)

I chose a Vector Autoregressive Model for several reasons. First, VAR models are fairly accurate when it comes to stationary, multivariate time series. Second, they are easy to interpret. Third, we can perform impulse response functions with a VAR to see if our theory aligns with the model's fit. A stationary VAR(p) model can be generalized to the following equation:

$$y_t = c + \Pi_1 y_{t-1} + \Pi_2 y_{t-2} + \Pi_2 y_{t-2} + \cdots + \Pi_n y_{t-p} + e_t \quad \textbf{(1)}$$

Π_i are nxn coefficient matrices and e_t is an $nx1$ unobservable white noise vector.[15]

The Vector Autoregressive Model in this paper predicts with horizons of 1 week, 1 month, and 3 months. Due to the recursive estimation required for forecasting, it is both costly and oftentimes inaccurate to use VARs for long range predictions. Thus, we only consider the VAR for relatively, short range forecasts. The order of the autoregressive component of the VAR was chosen by minimizing the Akaike Information Criteria (AIC).[16]

To ensure stationarity in the data, I perform Augmented Dickey-Fuller tests. We test against the null hypothesis that there is a unit root in our series. Basically, if we reject the null, then our data is probably stationary. Below is the result of the ADF Test:

Table 1. ADF Tests

ADF Test for: Unemployment Claims	ADF Test for: Welfare & Unemployment	ADF Test for: Job Search
ADF Statistic: -4.491722 P-value: 0.000204	ADF Statistic: -3.619092 p-value: 0.005406	ADF Statistic: -2.998136 p-value: 0.035069
Critical Values: 1%: -3.439 5%: -2.865 10%: -2.569	Critical Values: 1%: -3.439 5%: -2.865 10%: -2.569	Critical Values: 1%: -3.439 5%: -2.865 10%: -2.569

We reject the null in all cases. Thus, we may proceed with the assumption that our VAR is stationary. Looking at table 2, we can see the VAR output executed in levels. Note that AIC is minimized at p=5. Thus, we have a VAR(5) model. In the equation for claims, there is high significance for the first lags: *L1.claims*, *L1.welfare*. This agrees with the theory that Google searches can be a leading indicator of unemployment. I explore this further with a final step. I used impulse response functions, to further visualize the relationship estimated by the model fit. If we turn to figure 3, we can see that an increase in *Welfare & Unemployment* searches leads to an increase in unemployment claims. When people first search for unemployment benefits, it seems logical that they would soon after be considered unemployed. According to our model, Increases in *Job search* queries, would lead to a decrease in unemployment claims. Baker and Fradkin (2011) found job related searches increase when unemployment beneficiaries were

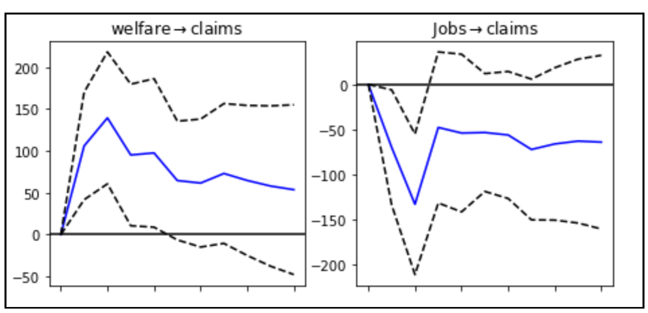
[15] https://faculty.washington.edu/ezivot/econ584/notes/varModels.pdf
[16] H. Akaike, "A new look at the statistical model identification," in IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716-723, December 1974, doi: 10.1109/TAC.1974.1100705.

within 10 weeks of their benefits ending.[17] Unemployment benefits lasted at most 53 weeks during the great recession.[18] As those unemployment benefits ended, the unemployment rate had started to go down, and so did the initial claims. In addition to this theory, it also makes sense that increases in people searching for jobs could show that the economy is recovering, and therefore, it would make sense that claims would go down as a result. Furthermore, claims go up with searches related to *welfare & unemployment* and claims go down with searches related to *job search* according to our VAR. Our model's fit to the data agrees with the theory.

Table 2. VAR(5) Summary

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Mon, 17, May, 2021
Time:                     23:51:33
--------------------------------------------------------------------
No. of Equations:         3.00000    BIC:                    30.7380
Nobs:                     680.000    HQIC:                   30.5424
Log likelihood:          -13189.0    FPE:                1.62454e+13
AIC:                      30.4188    Det(Omega_mle):     1.51506e+13
--------------------------------------------------------------------
Results for equation claims
=============================================================================
                coefficient       std. error           t-stat            prob
-----------------------------------------------------------------------------
const           3756.321600      5902.184612            0.636           0.524
L1.claims          0.701005         0.038286           18.310           0.000
L1.welfare       105.773341        32.725834            3.232           0.001
L1.Jobs          -69.853077        32.578519           -2.144           0.032
L2.claims         -0.000280         0.047071           -0.006           0.995
L2.welfare         7.706718        37.305014            0.207           0.836
L2.Jobs          -43.948647        38.445346           -1.143           0.253
L3.claims          0.167323         0.046522            3.597           0.000
L3.welfare       -45.617564        37.102690           -1.229           0.219
L3.Jobs          102.579856        38.508478            2.664           0.008
L4.claims         -0.022495         0.046975           -0.479           0.632
L4.welfare       -13.075885        37.276947           -0.351           0.726
L4.Jobs          -19.660429        38.910807           -0.505           0.613
L5.claims          0.149305         0.038199            3.909           0.000
L5.welfare       -53.144346        32.943516           -1.613           0.107
L5.Jobs           21.265535        32.802084            0.648           0.517
=============================================================================
```

Figure 3. Impulse Responses

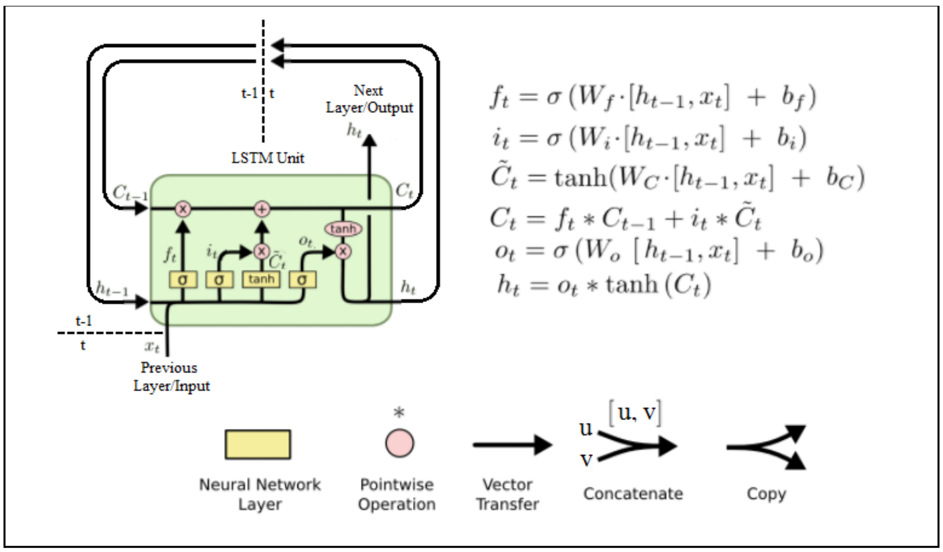


[17] Scott R. Baker and Andrey Fradkin, "What Drives Job Search? Evidence from Google Search Data," *SSRN Electronic Journal*, 2011, 3|, doi:10.2139/ssrn.1811247)
[18] Gary Burtless, "Unemployment Insurance for the Great Recession," Brookings, July 28, 2016, |PAGE|, accessed May 18, 2021, https://www.brookings.edu/testimonies/unemployment-insurance-for-the-great-recession/)

B. LSTM

Long Short-Term Memory networks were invented by Hochreiter and Schmidhuber in 1997 to solve the short-term dependency problem in Recurrent Neural Networks (RNN).[19] Previously, RNNs were good at solving problems that dealt with short-term memory. Formally, this problem is known as the vanishing gradient problem. LSTMs solve this by passing both a cell state and a hidden state between each cell. I will explain this further from figure 4.

Figure 4. LSTM Diagram



In this paper, I plan to explain the intuition behind why LSTM works, but I will admit it takes a fair amount of background knowledge in neural networks to understand everything. There are two main parts to seeing how information flows in this network. The first is the cell C_t and the hidden h_t states. These vectors are meant to store information to be passed to the next cell. The next parts are the three gates. First, the forget gate f_t which is meant to determine if any values in the cell state should be reset to zero. Second, there is the input gate i_t which is meant to determine if any values of the cell state should be updated. Finally, there is the output gate o_t which determines which information in the current hidden state should be passed on. Note, we use sigmoid functions for each of these three gates to produce a differentiable output between 0 and 1. This is important for when the neural network makes its backward pass to update the weights. The tanh layers redistribute the gradients to prevent vanishing or exploding gradients that are common in neural networks.

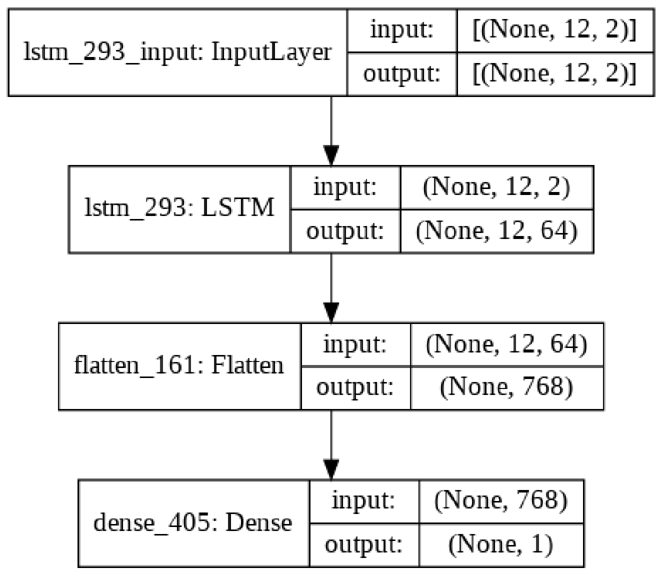
A more intuitive explanation of an LSTM is that the cell state holds all of the important long-term information whereas the hidden state holds all of the short-term information. This isn't always true, but at the very least, LSTM does hold a lot more long-term memory compared to vanilla RNNs. Another thing to note is that the information inside the LSTM isn't constrained to

[19] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long Short-Term Memory." *Neural Computation* 9, no. 8 (1997): 1735-780. doi:10.1162/neco.1997.9.8.1735.

linear coefficients. Instead, the entire network learns a non-linear surface through trial and error. Because of this, Long Short-Term Memory networks are listed under a class of algorithms known as black box algorithms. Essentially, we cannot easily interpret how the LSTM fits to data as there is a sequence of dependence among the input variables.[20] What we lose in interpretability, we gain in the network being able to learn relationships such as lagged variables without having to explicitly specify like we do with VAR.

The implementation used in this paper uses a simple LSTM which then gets flattened into a 1d vector. The output is returned through the final neuron in a dense layer.[21]

Figure 5.



VI. Results

For both the VAR and the LSTM, I fed data between January 14, 2006 and the first week of 2019. All of the forecasts below have been made on data that neither of the models have seen. According to Nelson and Plosser (1982), it is common for macroeconomic time-series to be represented as a random walk. If initial claims were truly a random walk, then the best univariate forecast for y_t would be y_{t-1}. It is from this equation that we establish our baseline model.[22]

Both the VAR and LSTM perform well against the baseline model. The VAR seems to be better at short term predictions compared to the LSTM which has a more relaxed fit to the data. This could be because I used L2 regularization which penalizes the LSTM for having weights with high variance. Table 3 shows the results of the short term forecasts between models. Looking at the figure 6, we can quickly notice a key difference between the LSTM and the VAR.

---

[20] Maryem Rhanoui et al., "Forecasting Financial Budget Time Series: ARIMA Random Walk vs LSTM Neural Network," *IAES International Journal of Artificial Intelligence (IJ-AI)* 8, no. 4 (2019): 4, doi:10.11591/ijai.v8.i4.pp317-327)

[21] https://keras.io/api/layers/core_layers/dense/

[22] "(PDF) Trends and Random Walks in Macroeconomic Time Series ...,"1-10, accessed May 18, 2021, https://www.researchgate.net/publication/256991483_Trends_and_random_walks_in_macroeconomic_time_series_A_reappraisal)

The VAR seems to closely mimic the baseline model. This could be because a VAR model is based on a linear combination of previous values unlike the non-linear LSTM. VAR outperformed the LSTM for a one week and one month prediction horizon. Beyond that the LSTM performs better with surprisingly good results. These results agree with Kundu and Singhania (2020) beyond the three month horizon. They find that LSTM works better than VAR in a one month prediction period which slightly differs from these results. A difference between this paper and theirs is that they forecast the monthly unemployment rate whereas this paper focuses on initial claims which may lead to slightly different results.[23]

Table 3.

Prediction Horizon	VAR (MAPE)	LSTM (MAPE)	Baseline (MAPE)
1 week	2.68	2.86	2.72
1 month	2.80	2.83	3.51
3 months	2.93	2.89	3.73

Table 4.

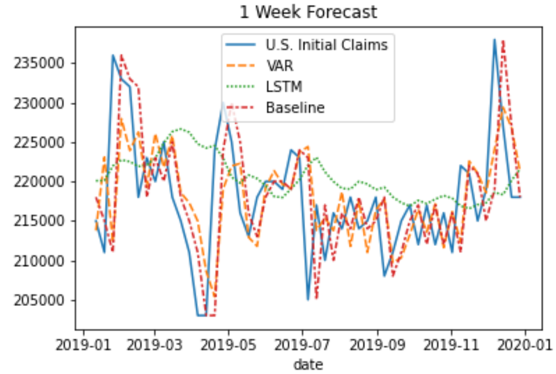
Prediction Horizon	LSTM (MAPE)	Baseline (MAPE)
6 months	3.38	3.92
1 year	3.20	3.57

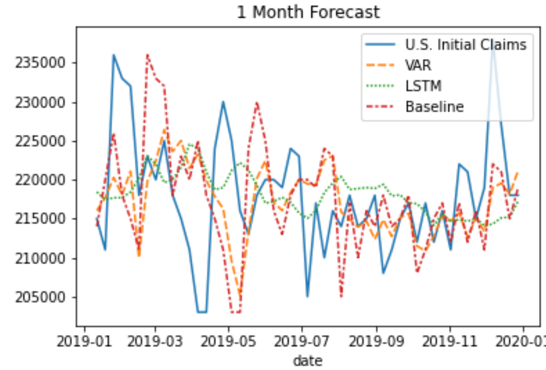
Another thing to note from the LSTM prediction was the hyperparameter tuning required to achieve the accuracy it did. I subsetted the training set into two to create a third validation set. From there, I fitted the LSTM as best as I could on the validation set. Sometimes, I had to deviate slightly from the architecture depicted in figure 5. Each pre-trained LSTM model will be made publicly available via the github repository for this project found here: [*https://github.com/jacobsomer/Google-Trends-and-Unemployment*](https://github.com/jacobsomer/Google-Trends-and-Unemployment). I will also publicly post the jupyter notebooks used to collect the data, execute the VAR, and create the LSTMs with their respective charts.

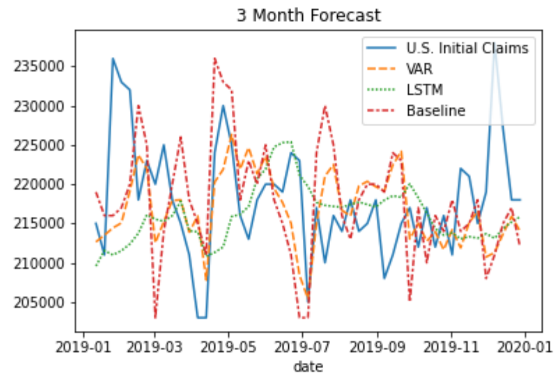
Figure 6. Forecasts

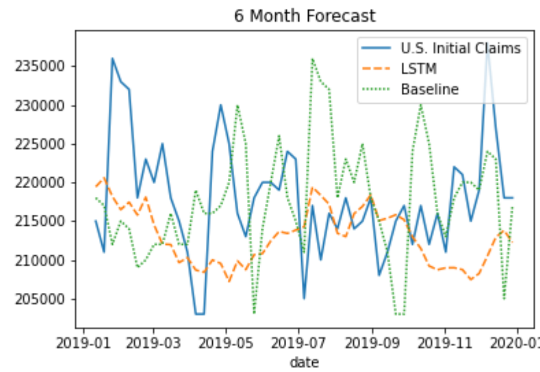
---

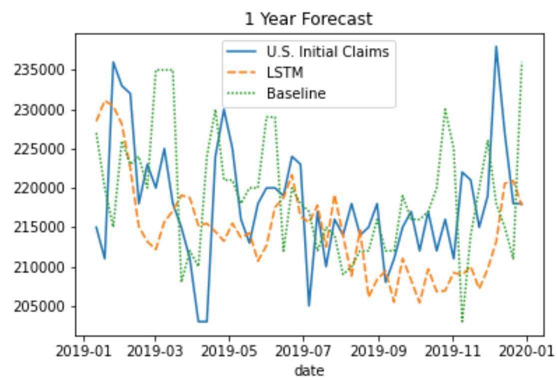
[23] ourav Kundu and Rajshekhar Singhania, "Forecasting the United States Unemployment Rate by Using Recurrent Neural Networks with Google Trends Data," *International Journal of Trade, Economics and Finance* 11, no. 6 (December 6, 2020): 1, doi:10.18178/ijtef.2020.11.6.679)











VII. Conclusion

Overall there is strong evidence that Google Trends predict U.S. Initial claims. The VAR estimates agreed with the economic theory behind the relationship between our keywords and weekly initial claims. When we compared models in forecasting, the VAR was good at predicting short-run time-series, but it was beaten by the LSTM for longer time periods. This is an interesting observation and could lead to promising results in economics/computing research moving forward. The main drawbacks from this study was the inability to interpret our LSTM model. Even though it made good long-term predictions, it is hard to say how those predictions were made. Luckily, previous research is on my side in that neural networks can perform better than traditional models as demonstrated by Maryem Rhanoui et al. Needless to say, we have statistically shown through our VAR estimate, that there is an underlying relationship between Google Trends and the Unemployment Rate. Moving forward, I seek to find ways to use internet data to detect anomalies. The covid spike in unemployment was really difficult to model, and I believe part of that was because I did not have the right keywords. Each major shock comes with its own unique set of keywords. As language changes, as websites come and go, there needs to be a better way to create indices for understanding economic phenomena in cyberspace.

VII. References

1. Askitas, Nikos, and Klaus F. Zimmermann. "Google Econometrics and Unemployment Forecasting." *SSRN Electronic Journal*, April 2009. doi:10.2139/ssrn.1480251.
2. Baker, Scott R., and Andrey Fradkin. "What Drives Job Search? Evidence from Google Search Data." *SSRN Electronic Journal*, 2011. doi:10.2139/ssrn.1811247.
3. Choi, Hyunyoung, and Hal Varian. "Predicting the Present with Google Trends." *Economic Record* 88 (June 2, 2012): 2-9. doi:10.1111/j.1475-4932.2012.00809.x.
4. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long Short-Term Memory." *Neural Computation* 9, no. 8 (1997): 1735-780. doi:10.1162/neco.1997.9.8.1735.
5. Kundu, Sourav, and Rajshekhar Singhania. "Forecasting the United States Unemployment Rate by Using Recurrent Neural Networks with Google Trends Data." *International Journal of Trade, Economics and Finance* 11, no. 6 (December 6, 2020): 135-40. doi:10.18178/ijtef.2020.11.6.679.
6. Rhanoui, Maryem, Siham Yousfi, Mounia Mikram, and Hajar Merizak. "Forecasting Financial Budget Time Series: ARIMA Random Walk vs LSTM Neural Network." *IAES International Journal of Artificial Intelligence (IJ-AI)* 8, no. 4 (2019): 317. doi:10.11591/ijai.v8.i4.pp317-327.
7. Rudebusch, Glenn D. "Trends and Random Walks in Macroeconomic Time Series: A Re-Examination." *International Economic Review* 33, no. 3 (1992): 661-80. Accessed May 18, 2021. doi:10.2307/2527132.
8. Varian, Hal R., and Hyunyoung Choi. "Predicting the Present with Google Trends." *SSRN Electronic Journal*, April 10, 2009. doi:10.2139/ssrn.1659302.
9. Werken, Laurens & Smit, Victoria. (2019). Exploring the Use of Recurrent Neural Networks for predicting Inflation and Unemployment. 10.13140/RG.2.2.29219.81446.