Using Google Search Volume to Predict Natural Gas Prices with Multiple LSTM Models

Quinn Murphey
University of Texas at San Antonio
1 UTSA Circle San Antonio, TX
quinn.murphey@my.utsa.edu

Adrian Ramos
University of Texas at San Antonio
1 UTSA Circle San Antonio, TX
adrian.ramos@my.utsa.edu

Gabriel Soliz
University of Texas at San Antonio
1 UTSA Circle San Antonio, TX
gabriel.soliz@my.utsa.edu

Abstract

The ability to accurately project the price of commodities is one of the most useful applications of deep learning. It finds use from hedge funds seeking to maximize profit to public administrations modelling the outcomes of different policies. In the typical year, these algorithms are quite successful, at least more so than their human counterparts. However, these models have almost always failed to predict drastic economic downturns such as the crash of oil in 2020 or the now expected crashes of Bitcoin. It's critical to everyone that we can prepare for sudden events that can drastically alter the markets. Building off of the work of Tang et al. [17], we use historical commodity prices along with Google search trends and news report sentimentatlity to hopefully achieve better commodity price predictions both in normal and abnormal times. In order to accomplish this task, we will use a combination of different deep learning algorithms including RNNs, CNNs, and GANs. We will compare our results with those from similar papers.

1. Introduction

The price fluctuation of good and stocks are often difficult to predict due to the numerous amounts of variables that play an important role of the price function. While there exists research that reflects on those expected variables [13]. Additionally, the research conducted which compares multiple results comprised from other researchers and their unique test leading to their results [15]. The importance of being able to accurately predict the price of commodities is vital to creating plans to aid those in need. The more accurate our forecasting ability is, the better prepared we can hope to be in uncertain times. It would allow emergency services and first responders to allocate enough supplies in the event of unpredictable events that could cause server damage to our infrastructure. However, there has been minimal research on the price fluctuation of goods and stocks due to external events, such as war, pandemics, or environmental catastrophes. While reports have been brought up that show certain effects of specific tragedies, such as the COVID-19 pandemic report [9]. The rate that prices fluctuate of goods and stocks during times of crisis and compared to other times of crisis could potentially help uncover areas which are most impacted. Including opportunities for potential preventive measures to attempt to thwart a severe effect.

The source code for our project can be found at https://www.github.org/Nragis/cs4263-project.

2. Related Work

From what we have observed there seems to be certain trends when trying to predict natural gas prices. The trend majority of the articles such as "Forecasting Natural Gas Spot Prices with Machine Learning" use is by taking the price of the gas as far as you have a data set for and then using adaptive and regression models to predict the gas prices future. The next theme that some articles use such as "Deep Neural Network Model for Improving Price Prediction of Natural Gas" is that they look at the current trend of natural gas and other similar items on something like google and if there is a trend of natural gas possibly becoming volatile with other forecasts also coming to this conclusion then it changes the prediction accordingly. The least common way that I have found is one explored in the paper "Natural Gas Price Prediction with Big Data" where the authors use sentiment analysis on a large body of literature, most commonly the news. This way while uncommon is surprisingly effective with it being able to tell the sentiment within the text and according to how drastic it is it changes the predictions.

3. Proposed Approach

For this project we will approach it in our own unique way. We will utilize the Energy Information Agency's Natural Gas dataset spanning the past several years. We will also utilize a time-series regression algorithm to analyze and predict the price for natural gas. Using a time-series regression algorithm should help us with utilizing and processing the data set we have chosen to its fullest extent utilizing every bit of knowledge we have to give an accurate prediction not only of the past but also the future. Utilizing this method our prediction data should be superior to the traditional econometric models and have the ability to predict future data points.

4. Data

4.1. Labels

To be able to compare directly with Tang et al. [17], we will use the same daily NYMEX natural gas futures prices from the US Energy Information Administration website (https://www.eia.gov/). These futures are for 1 month, 2 month, 3 month, and 4 month time periods. In alignment with Tang, we will be using data from these four contracts from January 2013 to June 2019. 1,638 records in total. Then, we will use simple linear interpolation to fill in any days without an entry like weekends or holidays. We end up with data from every day between January 2, 2013 to June 28, 2019, or 2,369 records total.

Figure 1. Data descriptions of NYMEX futures prices from Jan 2, 2013 to June 28, 2019.

	Mean	Std Dev	Skew	Kurtosis
Futures 1	3.172	0.718	0.637	0.202
Futures 2	3.302	0.683	0.501	-0.405
Futures 3	3.232	0.660	0.461	-0.540
Futures 4	3.249	0.637	0.491	-0.492

Finally, we take the logarithm of every data point to eliminate the exponential nature of financial data and standardize each column seperately using following equations for each element

$$x' = \frac{(x - \mu)}{\sigma} \tag{1}$$

where μ is the mean of the column and σ is the standard deviation.

Figure 2. Data descriptions of regularized NYMEX futures prices from Jan 2, 2013 to June 28, 2019

Regularized	Mean	Std Dev	Skew	Kurtosis
Futures 1	0	1.0	0.033	-0.087
Futures 2	0	1.0	0.026	-0.334
Futures 3	0	1.0	0.040	-0.475
Futures 4	0	1.0	0.091	-0.499

4.2. Features

4.2.1 NYMEX

We will be using historical data from the same NYMEX dataset mentioned above including the natural gas spot prices we did not use for our labels. However, we will nnow be using data ranging from January 5, 2004 to June 28, 2019, filling empty days using the same interpolation methods. We end up with 5,654 records of data, each with five points: one spot price and four futures prices. See Figure 3 for our graph of the NYMEX dataset.

We then regularize the data exactly as we did to our label data.

4.2.2 Google Trends

We will be using Google Trends (https://trends.google.com/) as our source for Google search history data. In this paper, we will be using the daily search volume data of 14 different terms, including "Natural Gas" from January 5, 2004 to June 28, 2019. We end up with the same number of 5,654 records as we saw in our NYMEX features. However, for our google trends dataset, we have 14 columns, one for each search term. See Figure 4 for data descriptive statistics and Figure 5 for each series plotted (by the month).

Each column is not directly comparable with the other columns, and is only comparable with itself. For example, if "Natural Gas" has 23 for a day while "Recession" has a 73, this does not mean that "Recession" was searched more that day. This only means that "Recession" was searched more on this day than a day where "Recession" is less than 73.

4.3. Formatting Data

Instead of our features consisting of one or N days prior to the day we're trying to predict (our label) using a time-series window. We will be using variable length time series features spanning from the beginning of our dataset to the day immediately prior to our label. This allows us to have as much information as possible, given our datasets, for each prediction.

The input to our model will look like $(Batch, Time, Features)$ where $Batch$ is fixed at

Figure 3. NYMEX Futures Prices from Jan 5, 2004 to June 28, 2019. Labels in orange and features in both blue and orange.

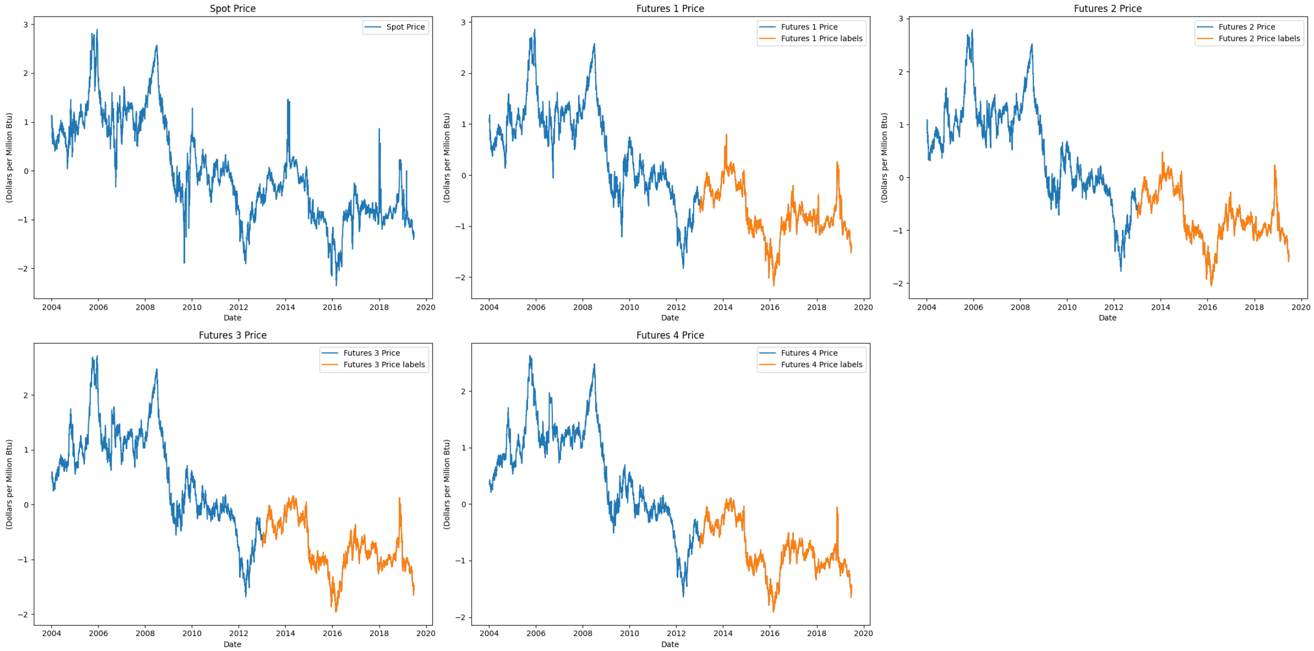

Figure 4. Data descriptions of daily Google search volume from Jan 5, 2004 to June 28, 2019.

	Mean	Std Dev	Skew	Kurtosis
Natural Gas	51.95	10.62	0.181	0.758
Oil	44.42	0.683	0.374	-0.898
Coal	24.19	0.660	0.368	0.276
Nuclear Power	5.662	4.064	7.544	135.1
Wind Power	20.34	13.15	1.396	2.496
Hydroelectric	15.68	12.59	0.852	0.424
Solar Power	35.28	12.69	0.676	0.697
Gold	40.15	13.01	-0.040	0.627
Silver	47.10	10.54	-0.224	0.176
Platinum	43.51	8.671	0.325	1.355
Copper	58.34	12.92	0.015	-0.237
Biofuel	12.82	12.15	1.762	4.112
Recession	5.728	6.258	3.348	16.91
CPI	20.55	11.41	1.031	1.893

16, $Time$ is variable, and $Feature$ is fixed at 19 (5 NYMEX and 14 Google).

The output of our model will be fixed however, looking like $(Batch, 1, Labels)$ where $Batch$ is 16, and $Labels$ is 4 (NYMEX Futures).

5. Metrics

In alignment with Tang, we will use mean absolute error (MAE) and root mean square error (RMSE) to compare results. Our goal is to minimize these values, indicating a more accurate regression.

$$MAE = \frac{1}{N}\sum_{i}^{N}|y_i - \hat{y}_i| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2} \tag{3}$$

Where y_i and $\hat{y}_i$ are the real and predicted values respectively.

6. Results

We used Python3 for all of our code. Additionally, we use Pandas [12] for our data fetch and preprocessing stages of our research, and we used Keras [5], and Tensorflow [1] for creating, training, and testing our models. Finally, we used Matplotlib [6] to create all of our plots.

We experimented with both stacked LSTMs and stacked BiLSTMs followed by several layers of densely connected perceptrons, and finally an output layer. To find the best model for our data, we performed a Bayesian hyperparameter optimization using Keras Tuner [11] over the hyperparameter dimensions listed in Figure 6).

7. Conclusion

We will write this section once we have final results.

Figure 5. Monthly Google search volume from Jan 5, 2004 to June 28, 2019

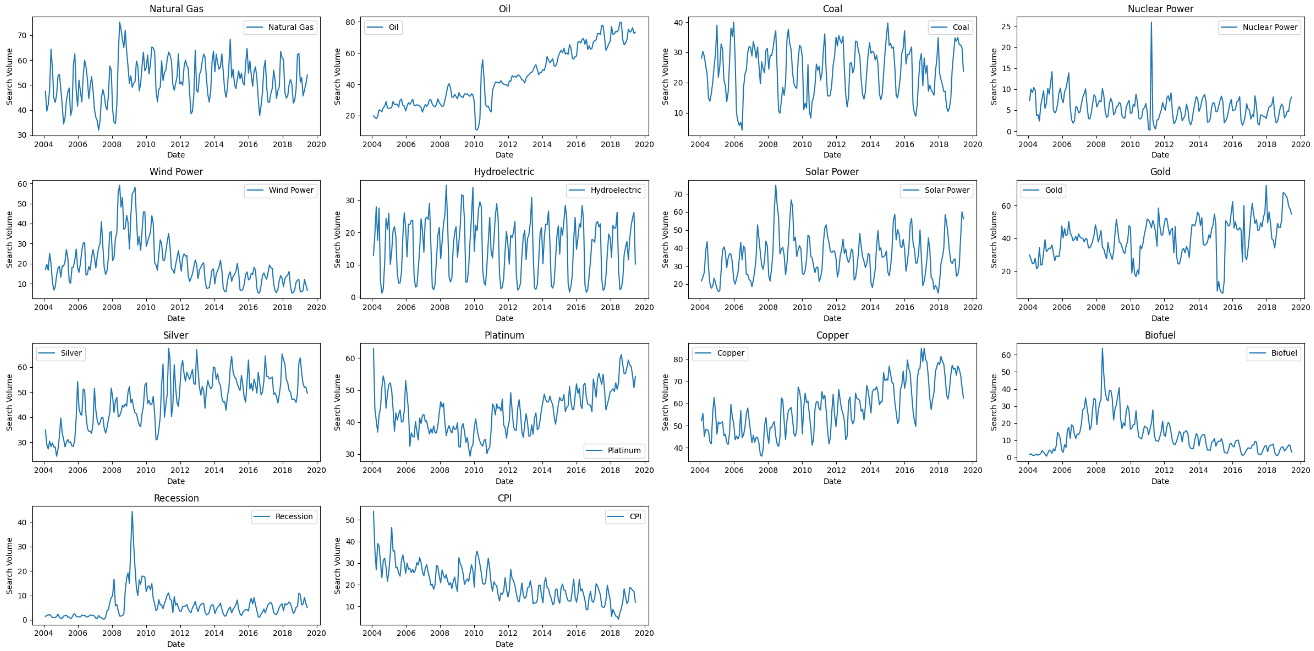

Figure 6. Hyperparameter distrbutions searched for our models

Name	Type	Min	Max	Distribution(Step)
LSTM Layers	Int	1	5	Uniform(1)
LSTM Nodes	Int	32	256	Uniform(32)
Dense Layers	Int	1	3	Uniform(1)
Dense Nodes	Int	256	2048	Uniform(256)
Dropout Rate	Float	0	0.999	Uniform
Learning Rate	Float	1e-6	1e-1	Log
Beta_1	Float	0.8	0.999	Linear

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 3

[2] Aliyuda Ali, M. K. Ahmed, Kachalla Aliyuda, and Abdulwahab Muhammed Bello. Deep neural network model for improving price prediction of natural gas. In *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, pages 113–117, 2021.

[3] Minh Triet Chau, Diego Esteves, and Jens Lehmann. A neural-based model to predict the future natural gas market price through open-domain event extraction. In *CLEOPATRA 2020, Cross-lingual Event-centric Open Analytics*, 2020.

[4] Jae Young Choi and Bumshik Lee. Combining lstm network ensemble via adaptive weighting for improved time series forecasting. *Mathematical Problems in Engineering*, 2018.

[5] Francois Chollet et al. Keras, 2015. 3

[6] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. 3

[7] Iris Kesternich, Bettina Siflinger, James P. Smith, and Joachim K. Winter. The effects of world war ii on economic and health outcomes across europe. *Institute for the Study of Labor (IZA), Research Paper Series*, (6296), 2012.

[8] Praveen Kumar, Priyanka Sihag, Pratik Chaturvedi, K.V. Uday, and Varun Dutt. Bs-lstm: An ensemble recurrent approach to forecasting soil movements in the real world. *Frontiers in Earth Science*, 9, 2021.

[9] Dave Mead, Karen Ransom, Stephen B. Reed, and Scott Sager. The impact of the covid-19 pandemic on the food price indexes and data collection. *Monthly Labor Review. U.S. Dept. of Labor, Bureau of Labor Statistics*, August 2020. 1

[10] Dimitrios Mouchtaris, Emmanouil Sofianos, Periklis Gogas, and Theophilos Papadimitriou. Forecasting natural gas spot prices with machine learning. *Energies*, 14(18), 2021. 5782.

[11] Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. Keras Tuner.

https://github.com/keras-team/keras-tuner, 2019. 3

[12] The pandas development team. pandas-dev/pandas: Pandas, Feb. 2020. 3

[13] Ricardo Alberto Carrillo Romero. Generative adversarial network for stock market price prediction, 2019. Stanford University CS230 Final Project. 1

[14] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. The performance of lstm and bilstm in forecasting time series. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 3285–3292, 2019.

[15] Sarvagya Srivastava, Vishwaas Khare, and R. Vidhya. Economic forecasting using generative adversarial networks. *International Journal of Engineering Research & Technology*, 10(5), 2021. 1

[16] Moting Su, Zongyi Zhang, Ye Zhu, Donglan Zha, and Wenying Wen. Data driven natural gas spot price prediction models using machine learning methods. *Energies*, 12(9), 2019.

[17] Yuanyuan Tang, Qingmei Wang, Wei Xu, Mingming Wang, and Zhaowei Wang. Natural gas price prediction with big data. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5326–5330, 2019. 1, 2

[18] Yupeng Wang, Shibing Zhu, and Changqing Li. Research on multistep time series prediction based on lstm. In *2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)*, pages 1155–1159, 2019.