Using Google Search Volume and News Sentiment to Predict Natural Gas Prices with LSTMs

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. [7], 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 [5]. Additionally, the research conducted which compares multiple results comprised from other researchers and their unique test leading to their results [6]. 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 [3]. 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 senti-

ment 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. Commodity Prices

To be able to compare directly with Tang et al. [7], 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.

In future updates, we will have more papers with different types of commodities, specifically those not related to energy. Until we get better results with the NYMEX dataset, we will not be working with other data.

4.2. Internet Search History

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 for the respective commodity: natural gas, [COMMODITY2], and [COMMODITY3] and a few recession related terms such as [RECESSION_TERMS] each covering the exact time period of the commodity price dataset. These datasets are 2,372, [BLANK], and [BLANK] records long for natural gas, [COMMODITY2], and [COMMODITY3] respectively.

4.3. News Report Sentimentality

We collected the title and bodytext of all news articles from Yahoo Finance (We want to experiment with different news sources, both financial and not, and both credible and not) with the keyword natural gas, [COMMODITY2], and [COMMODITY3] respectively, each covering the exact time period of the commodity price dataset. We do not have an exact length for this dataset due to still deciding which news sites to use.

5. Experiments

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.

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

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

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

6. Results

Note: Our results relate to a prior direction of our paper, meaning they do not use the same datasets as we mention above.

Using the World Economic Overview dataset published by the International Monetary Fund semi-annually which tracks 44 economic indicators including Gross Domestic Product (GDP), Gross Domestic Product per Capita (GDPPC), and Average Consumer Prices Inflation Index (PCPI) for 194 countries from 1980 to 2020. While the database isn't completely full, most of the entries of the 8820 row, 57 column dataset are filled.

After which, we turn the dataset into a multivariable (each indicator) time series for each country. We then pass a 10 width window over each time series, creating a data set of 6208 rows - each with a country, 396 input variables (44 indicators across 9 years) and one output variable: the tenth year value of GDPPC (in purchasing power).

We then split the dataset into a 0.8, 0.1, 0.1 split for training, validation, and test datasets.

We compared four basic models to each other:

1. Linear: A simple matrix multiplication represented by a single layer, single perceptron neural network.

2. Single-Layer Perceptron: A single layer of 512 neurons with ReLU acivation functions.

3. Multi-Layer Perceptron: Two layers of 512 neurons with ReLU activation functions

4. CNN Without Pooling: Three 1D convolutional layers with 3 length kernels and 256, 128, and 64 filters

respectively followed by one layer of 256 ReLU perceptrons.

Our results were as follows:

TABLE 1: COMPARISON OF GDPPC PREDICTION RESULTS ACROSS DIFFERENT MODELS

Model	MAE
Linear	379.37
Single	20.31
Multi	19.74
CNN	16.00

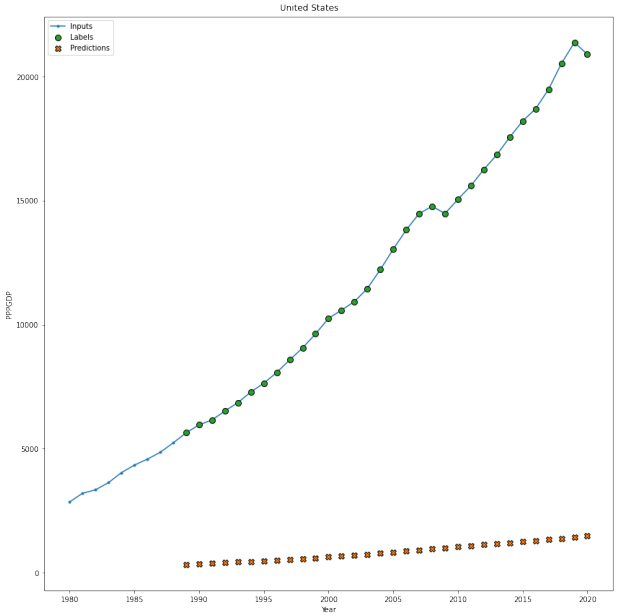

FIGURE 2: LINEAR MODEL PREDICTIONS OF GDPPC

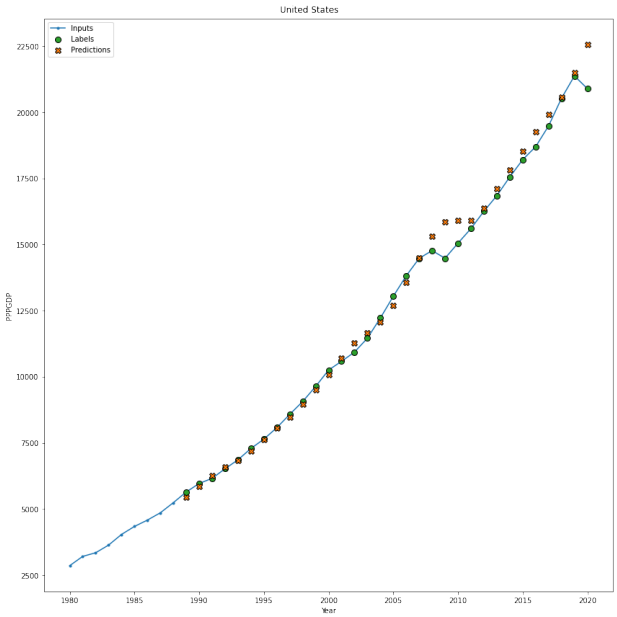

FIGURE 3: SINGLE-LAYER PERCEPTRON MODEL PREDICTIONS OF GDPPC

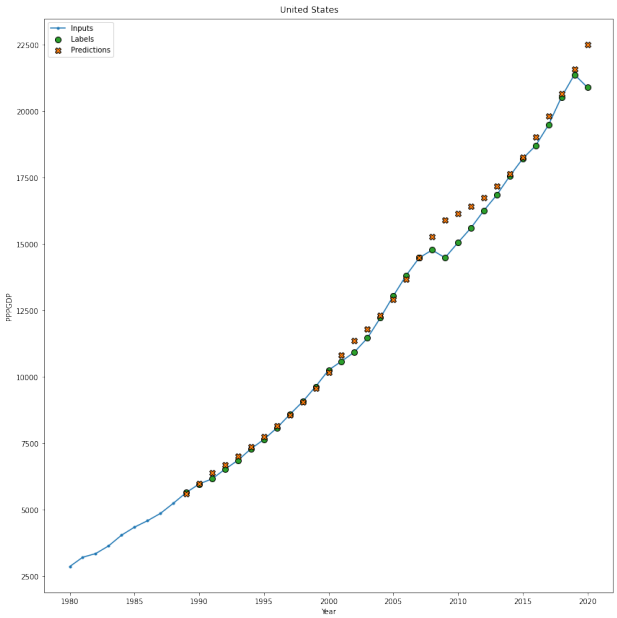

FIGURE 4: MULTI-LAYER PERCEPTRON MODEL PREDICTIONS OF GDPPC

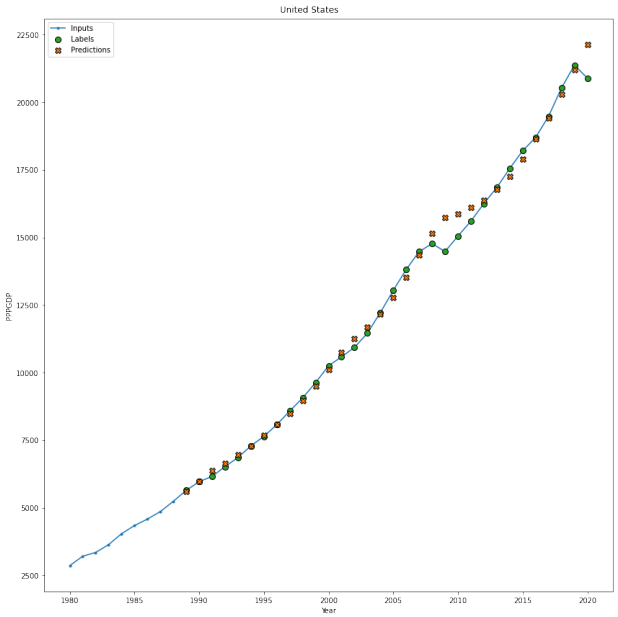

FIGURE 5: CNN WITHOUT POOLING MODEL PREDICTIONS OF GDPPC

7. Conclusion

We will write this section once we have final results.

References

[1] 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.

[2] 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.

[3] 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

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

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

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

[7] 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