The US Inflation Phenomenon | It’s Oil, silly

Author | Rand Sobczak Jr.

Date | 21 April 2021

Table of Contents

1.0 Problem Identification Overview
2.0 Generated Deliverables
3.0 Data Pre-Processing Steps
4.0 Model Description
5.0 Model Findings
6.0 Next Steps

1.0 Problem Identification Overview

The United States Consumer Price Index (**"Inflation"**) is calculated by the U.S. Bureau of Labor Statistics. It has gone through various periods of prominent increases; notably in the 1920's, 1940's, & 1970's. Otherwise, it has remained relatively constant or declining.

Inflation is important in all facets of life but the financial world pays special attention to it as the key objectives of the Federal Reserve are maximizing employment, stabilizing prices & moderating long-term interest rates; the second of which is Inflation & the third of which is generally decided by the condition of the other two (2), usually Inflation, & can move financial markets around the world; we won't go into that here. In short, **Inflation is an important component of developing investment strategies for portfolios across the world**. **The view on inflation becoming positive or negative is not agreed upon nor are the variables which influence it**.

The purpose of this Data Science project is to develop a model to explain & understand the phenomenon of Inflation. **I shortlisted nineteen (19) variables ("Variables") to determine their influence on Inflation found in Appendix I.**

2.0 Generated Deliverables

The following Application Programming Interface's (API) were used to pull the relevant data:

1. Quandl
2. Investing.com ("**Investpy**")
3. Federal Reserve Economic Data ("**FRED**")

The data retrieved by the API's & respective websites **are from sources highlighted in Appendix I**.

With this, **three (3) deliverables were generated**:

1. The source code for the modeling developed to analyze the aforementioned problem (link)
2. This document outlining the process
3. A PDF presentation with our findings (link)

3.0 Data Pre-Processing Steps

Data Cleaning:

1. The data for the Variables & Inflation were first pulled using the API's
2. The result, however, produced a data frame with non-congruent lengths in time (**Appendix II**)
3. I then was required to draw the line at the 4 April 1990 to ensure they aligned appropriately (**Appendix II**)
4. The resulting data frame was fully comprised of 9,616 non-null float values
5. I concatenated the Variables with Inflation using a Forward Fill technique as their reporting schedules did not align. This resulted in a Daily (Mon-Fri (excluding holidays)) data frame with only 317 observations (**Appendix II**)
6. This data frame was used to cross reference the accuracy of the next steps in Exploratory Data Analysis

Exploratory Data Analysis:

I was unsure as to which time periods for percent changes to use for the Variables & Inflation. The first step was to **run through the process on multiple configurations** as per below:

1. **Quarter-on-Quarter**
   a. **Description** | This looked at Quarterly percent change in both Variables & Inflation
   b. **Result** | **Reasonable**
2. **Month-on-Month**
   a. **Description** | This looked at Monthly percent change in both Variables & Inflation
   b. **Result** | **Poor**
3. **Quarter-on-Quarter for Variables (past) & Inflation (forwards)**
   a. **Description** | This looked at Quarterly percent change in Variables looking backwards while a forward looking Quarterly change in Inflation to ascertain if changes in the Variables took time to reach Inflation
   b. **Result** | **Poor**
4. **Quarter-on-Quarter w/ Rolling Averages on Daily, Weekly & Monthly Variables ("Best Configuration")**
   a. **Description** | This approach is similar to # 1 (looking at the Quarterly percent change in both Variables & Inflation) albeit used a rolling average for those that were reported more often than once a Quarter. The rational was that a Variable may have had a bad week or day when the Quarter ended; as such, the entire changes throughout the Quarter may need to be accounted for evenly
   b. **Result** | **Best**

The respective **Feature Correlation Heat Maps with the Pearson correlation coefficients are found in Appendix III** which will go into greater detail for the rankings listed in the Results (b) above.

It should be important to note that Inflation was scraped ±3% on all four (4). I went further to scrape Variables as well on #4 above albeit those results became even worse; you can find those results at the end of my source code (link) after 1.4.3 therein if you wish.

In summary, **the Best Configuration** (**#4**) **was chosen as it presented the most optimal Pearson correlation coefficients to achieve our goal of developing a model to explain & understand the phenomenon of Inflation.**

Pre-processing:

1. The first step in pre-processing, per the Machine Learning process, was to create a "Best Guess" number, with the help of a standard mean & the DummyRegressor functions. The data frame has no categorical data so this step was purely to "go through the process" but irrelevant; **please disregard this step when reading the source code**
2. I then went ahead & **split the data into training and testing splits, 70% and 30% respectively**
3. Next was to **scale the data**. **I wasn't sure which scaling technique to use** & thus, **I applied the following on both the X & y variables & X only** (x6 in total):
   a. Standard Scaling (SS)
   b. MinMax Scaling (MM)
   c. Log Transformation (LG)
4. As all data have their unique structures, **I then divided the Variables into selected groups to apply SS & LG separately to where SS & LG may be better suited**; this was merged to an unscaled y (Inflation)
5. The **initial results** (**Appendix IV**) showed that MM presented poor results for R^2, Mean Absolute Error (MAE), & MSE; thus was taken out of consideration
6. I now had x5 scaling approaches used & determined that all of the five (5) showed results that should be put to the process of the Random Forest Generator to identify the best Variables (**Appendix V**)

4.0 Model Description

Due to its better performance, **the model I used was the Random Forest Model with the goal of determining what variables best explain & understand Inflation**.

5.0 Model Findings

As the purpose of this Data Science project is to develop a model to explain & understand the phenomenon of Inflation, **I went through the following process on the five (5) shortlisted scaling approaches**:

- Grid Search
- Random Forest
- Hyperparameter search using Grid Search CV

The final outcomes can be seen in **Appendix VI**.

Wages CPI held a ubiquitous position as being the dominate Variable on all scaling approaches. **This is justified due to it's connection with how Inflation is calculated within**. Wages CPI may also summarize Initial Jobless Claims & Unemployment Rate leaving them potentially redundant.

WTI held second place on all scaling approaches. **I believe this makes sense given it's incorporated in just about every activity & purchase made in the United States**. Take for example a loaf of bread. The grains required to make dough are usually plowed by a tractor using gasoline (a derivative of oil). It is then transported to a bakery (using a derivative of oil). It's then transported to a store, purchased by a customer & driven home (again, using a derivative of oil).

Due to their dominance on all scaling approaches, **I then removed the other Variables & ran the process with only Wages CPI & WTI. The results are below**:

```
R² results for X & y scaled below
SS Train | 0.2924    Test 0.424
LG Train | 0.2815    Test 0.3673

R² results for X only scaled below
SS Train | 0.2924    Test 0.3489
LG Train | 0.2778    Test 0.2979

R² results for the LG & SS combination below
SS Train | 0.284    Test 0.3761
```

```
MAE results for X & y scaled below
SS Train | 0.5639    Test 0.5811
LG Train | 0.5727    Test 0.598

MAE results for X only scaled below
SS Train | 0.4515    Test 0.6127
LG Train | 0.4572    Test 0.6272

MAE results for the LG & SS combination below
SS Train | 0.4572    Test 0.6272
```

```
MSE results for X & y scaled below
SS Train | 0.7076    Test 0.6877
LG Train | 0.7185    Test 0.7763

MSE results for X only scaled below
SS Train | 0.4536    Test 0.7774
LG Train | 0.4629    Test 0.8615

MSE results for the LG & SS combination below
SS Train | 0.4629    Test 0.8615
```

After review of the **Test set results** (**the arbiter**), the **SS on both X & y was chosen** amongst the other scaling approaches **given it had the highest R^2 & lowest MAE & MSE**.

So **how do these test results compare to those seen after rolling averages on the unscaled 19 Variables**? (below)

```
A 17.0 bps increase in R²; 66.94 % increase.

A 1.81 bps increase in MAE.

A -6.79 bps decrease in MSE.
```

Further to this, **how do these test results compare to those seen after rolling averages on the SS X & y scaled 19 Variables** which was the best of the scaling approaches? (below)

```
A 14.44 bps increase in R²; 51.64 % increase.

A -10.29 bps decrease in MAE.

A -17.24 bps decrease in MSE.
```

Our conclusion, outside of Wages CPI (a component of Inflation itself), **how to explain & understand the phenomenon of Inflation** is a reconfiguration of words used by James Carville (link) | **It's Oil, silly**.

6.0 Next Steps

Inflation is a difficult & highly disputed financial beast but the closer you get to taming it your eyes will open wider.

Throughout the process there were a number of things I either wanted to change or add. **I will start with some Variables which were not included** due to financial & time constraints:

- Steel
    - I could only get it back to 2008
- Gasoline
    - I could only get it back to 2005
- Growth of M2
    - I put it aside to mitigate any overlap with M2 Velocity
- US Wages Hourly Earnings
    - Limited data as well
- US Dollar Index: Broad, Goods and Services
    - Only goes back to 2006 (discussed below)

Further to these, **further consideration may be applicable to the below**:

- Reassess the Variables which were chosen in the SS & LG divide; discussed in Pre-processing #5 above
- Although scraping on Variables was investigated (see “Scraping the Variables as well (individually)” in the Data Wrangling & EDA link here), a pivot to Winsorizing way present better results
- Develop a model to predict Wages CPI itself in order to remove ourselves from the US gov’t’s reporting
- I believe the US Dollar Index Variable (DXY) does not correctly address the situation of the United States either Imports or Exports Inflation (usually the former). This may be because the DXY’s weighting does not correctly align to the US’s trade. In short, it’s a weighted geometric mean of the Eurozone (EUR), Japan (JPY), the United Kingdom (GBP), Canada (CAD), Sweden (SEK) & Switzerland (CHF) but that does not correctly align with the US’s trade with the world

There are a number of Variables to take into consideration but **for now I would recommend keeping an eye on Oil.**

Appendix I

Items	Reported	API	API Source	Comments
Inflation	Monthly	Quandl	U.S. Bureau of Labor Statistics	The target variable
Wages CPI	Monthly	FRED	U.S. Bureau of Labor Statistics	A component of the target variable
WTI	Daily	Quandl	CME	West Texas Intermediate - One of many commodities
Heating Oil	Daily	Investpy	Investing.com	One of many commodities
Copper	Daily	Investpy	Investing.com	One of many commodities
Sugar	Daily	Investpy	Investing.com	One of many commodities
Natural Gas	Daily	Investpy	Investing.com	One of many commodities
Cattle	Daily	Investpy	Investing.com	One of many commodities
Lean Hogs	Daily	Investpy	Investing.com	One of many commodities
Soybeans	Daily	Investpy	Investing.com	One of many commodities
Lumber	Daily	Investpy	Investing.com	One of many commodities
Capacity Utilization	Monthly	FRED	Board of Governors of the Federal Reserve	The % of resources used by corporations
Corn	Daily	Investpy	Investing.com	One of many commodities
M2 Velocity	Quarterly	FRED	Federal Reserve Bank of St. Louis	Movement of money; state of the economy proxy
GDP	Quarterly	FRED	U.S. Bureau of Economic Analysis	A proxy for the state of the economy
Wheat	Daily	Investpy	Investing.com	One of many commodities
PMI	Monthly	Quandl	Institute of Supply Management	Manufacturing PMI - A proxy for the economy
USD Index	Daily	Quandl	Intercontinental Exchange Inc	(DXY) Proxy for potentially importing inflation
Unemployment Rate	Monthly	Quandl	U.S. Bureau of Labor Statistics	A proxy for the state of the economy
Initial Jobless Claims	Weekly	Quandl	U.S. Employment and Training Administration	A proxy for the state of the economy

Appendix II

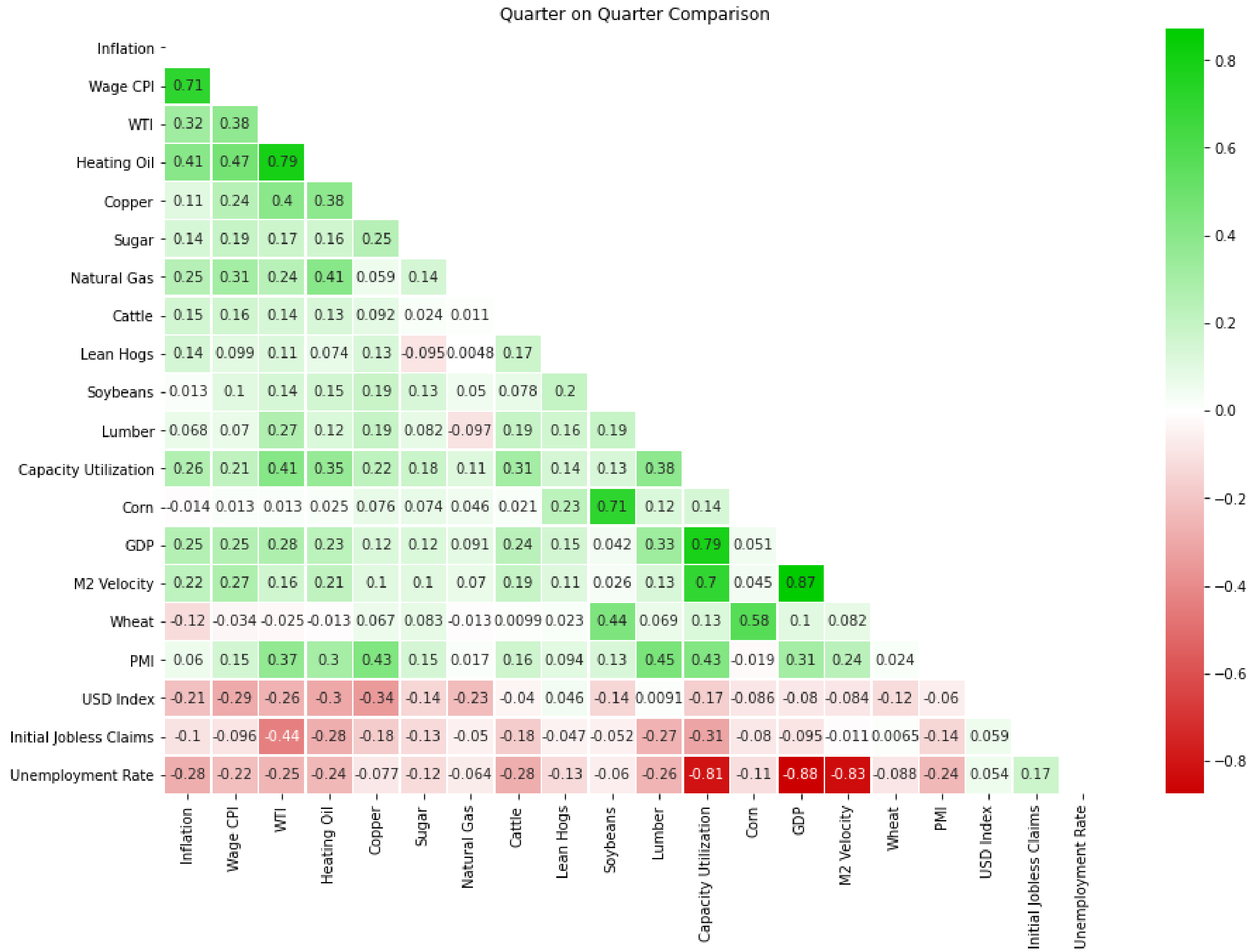
```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 14166 entries, 1946-01-01 to 2021-04-20
Data columns (total 19 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Wages CPI               14157 non-null  float64
 1   WTI                     11952 non-null  float64
 2   Copper                  10304 non-null  float64
 3   Soybeans                9863 non-null   float64
 4   Natural Gas             9779 non-null   float64
 5   Heating Oil             12951 non-null  float64
 6   Corn                    12950 non-null  float64
 7   Wheat                   9865 non-null   float64
 8   Cattle                  12948 non-null  float64
 9   Lean Hogs               12953 non-null  float64
 10  Sugar                   12951 non-null  float64
 11  Lumber                  12953 non-null  float64
 12  Capacity Utilization    13897 non-null  float64
 13  GDP                     14159 non-null  float64
 14  M2 Velocity             14015 non-null  float64
 15  PMI                     14145 non-null  float64
 16  USD Index               11137 non-null  float64
 17  Initial Jobless Claims  13894 non-null  float64
 18  Unemployment Rate       14145 non-null  float64
dtypes: float64(19)
memory usage: 2.2 MB
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 9616 entries, 1991-04-18 to 2021-04-20
Data columns (total 19 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Wages CPI               9616 non-null   float64
 1   WTI                     9616 non-null   float64
 2   Copper                  9616 non-null   float64
 3   Soybeans                9616 non-null   float64
 4   Natural Gas             9616 non-null   float64
 5   Heating Oil             9616 non-null   float64
 6   Corn                    9616 non-null   float64
 7   Wheat                   9616 non-null   float64
 8   Cattle                  9616 non-null   float64
 9   Lean Hogs               9616 non-null   float64
 10  Sugar                   9616 non-null   float64
 11  Lumber                  9616 non-null   float64
 12  Capacity Utilization    9616 non-null   float64
 13  GDP                     9616 non-null   float64
 14  M2 Velocity             9616 non-null   float64
 15  PMI                     9616 non-null   float64
 16  USD Index               9616 non-null   float64
 17  Initial Jobless Claims  9616 non-null   float64
 18  Unemployment Rate       9616 non-null   float64
dtypes: float64(19)
memory usage: 1.5 MB
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 317 entries, 1991-04-30 to 2021-03-31
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Inflation               317 non-null    float64
 1   Wages CPI               317 non-null    float64
 2   WTI                     317 non-null    float64
 3   Copper                  317 non-null    float64
 4   Soybeans                317 non-null    float64
 5   Natural Gas             317 non-null    float64
 6   Heating Oil             317 non-null    float64
 7   Corn                    317 non-null    float64
 8   Wheat                   317 non-null    float64
 9   Cattle                  317 non-null    float64
 10  Lean Hogs               317 non-null    float64
 11  Sugar                   317 non-null    float64
 12  Lumber                  317 non-null    float64
 13  Capacity Utilization    317 non-null    float64
 14  GDP                     317 non-null    float64
 15  M2 Velocity             317 non-null    float64
 16  PMI                     317 non-null    float64
 17  USD Index               317 non-null    float64
 18  Initial Jobless Claims  317 non-null    float64
 19  Unemployment Rate       317 non-null    float64
dtypes: float64(20)
memory usage: 52.0 KB
```

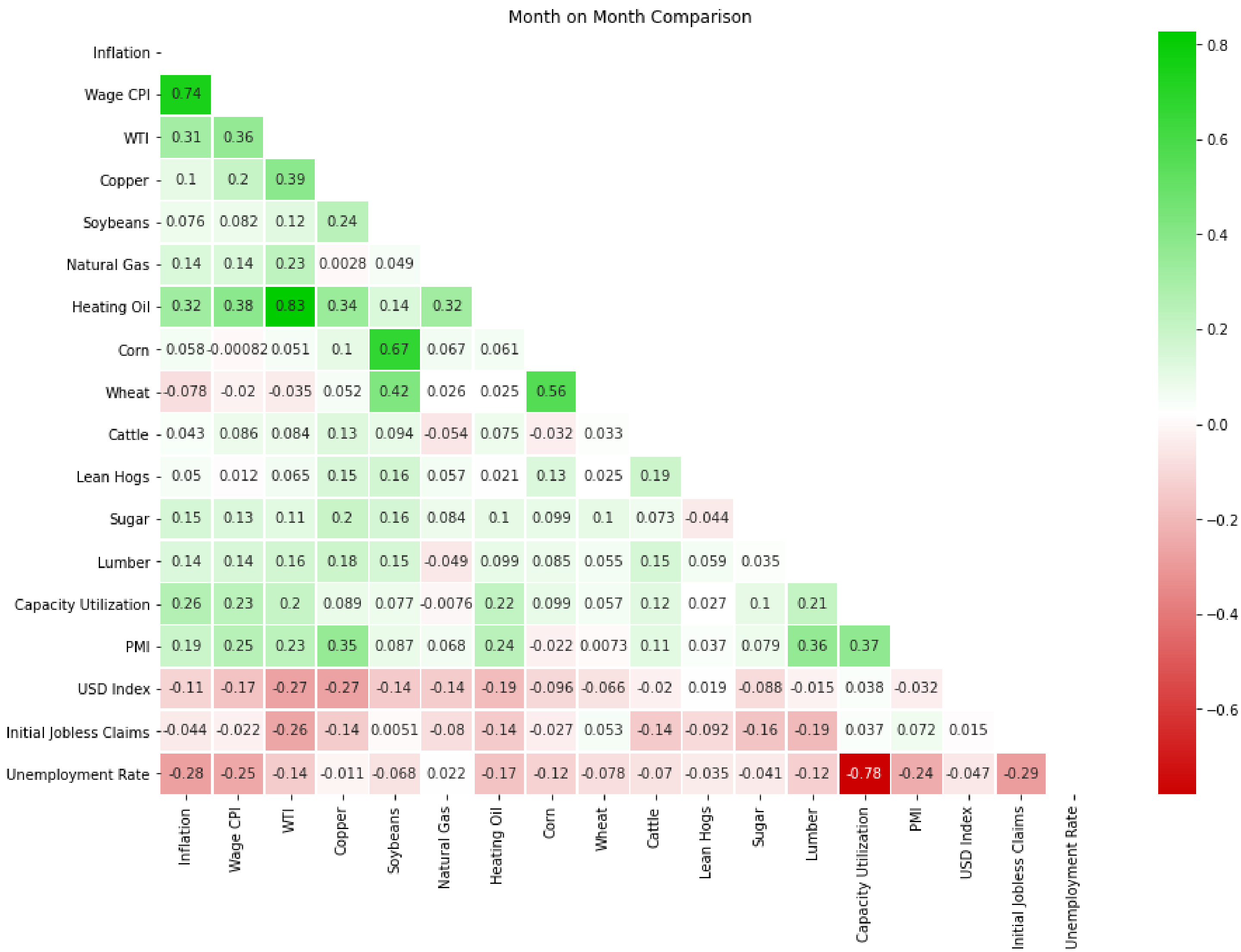
Appendix III (Feature Correlation Heatmap)

#1



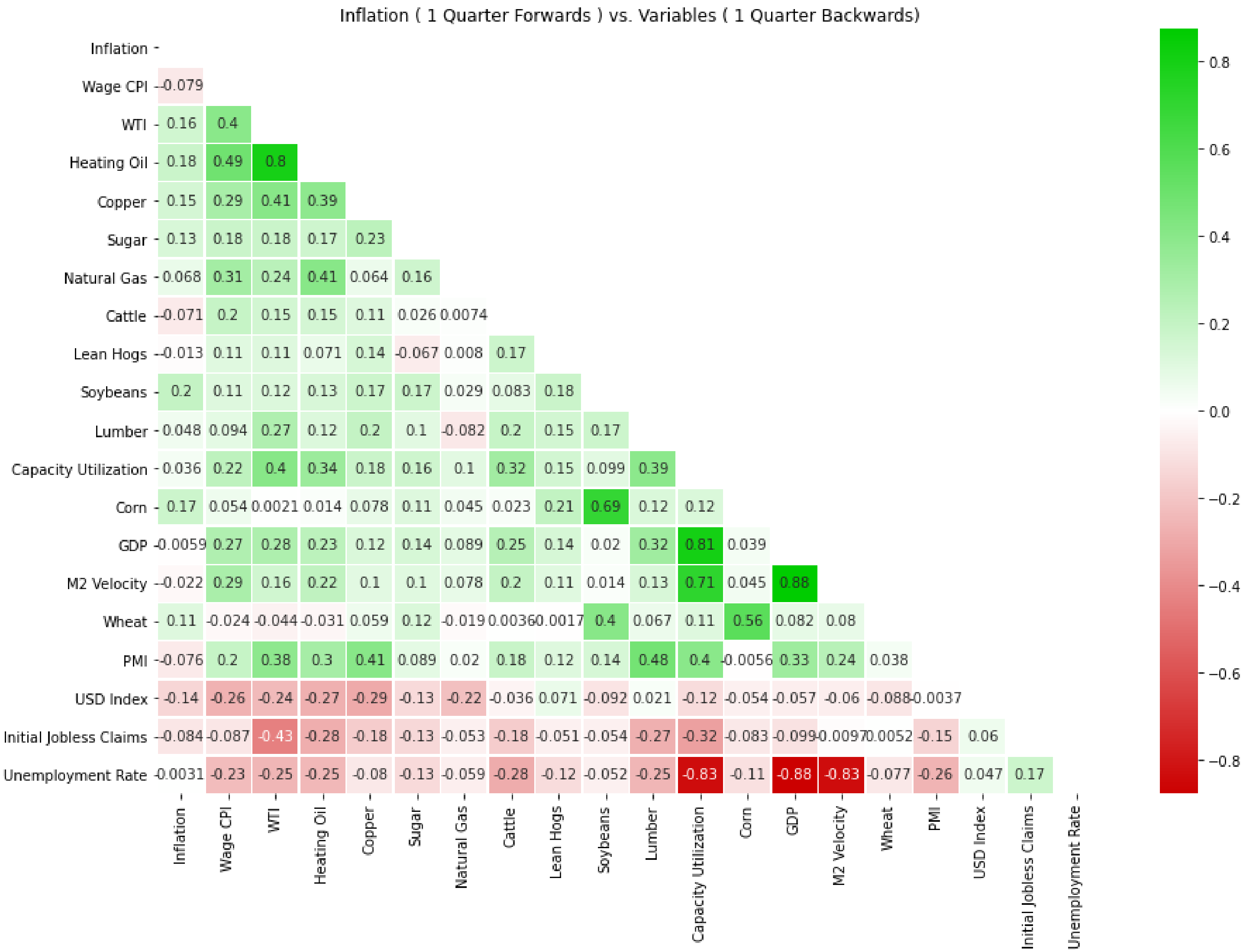
Appendix III (Feature Correlation Heatmap)

#2



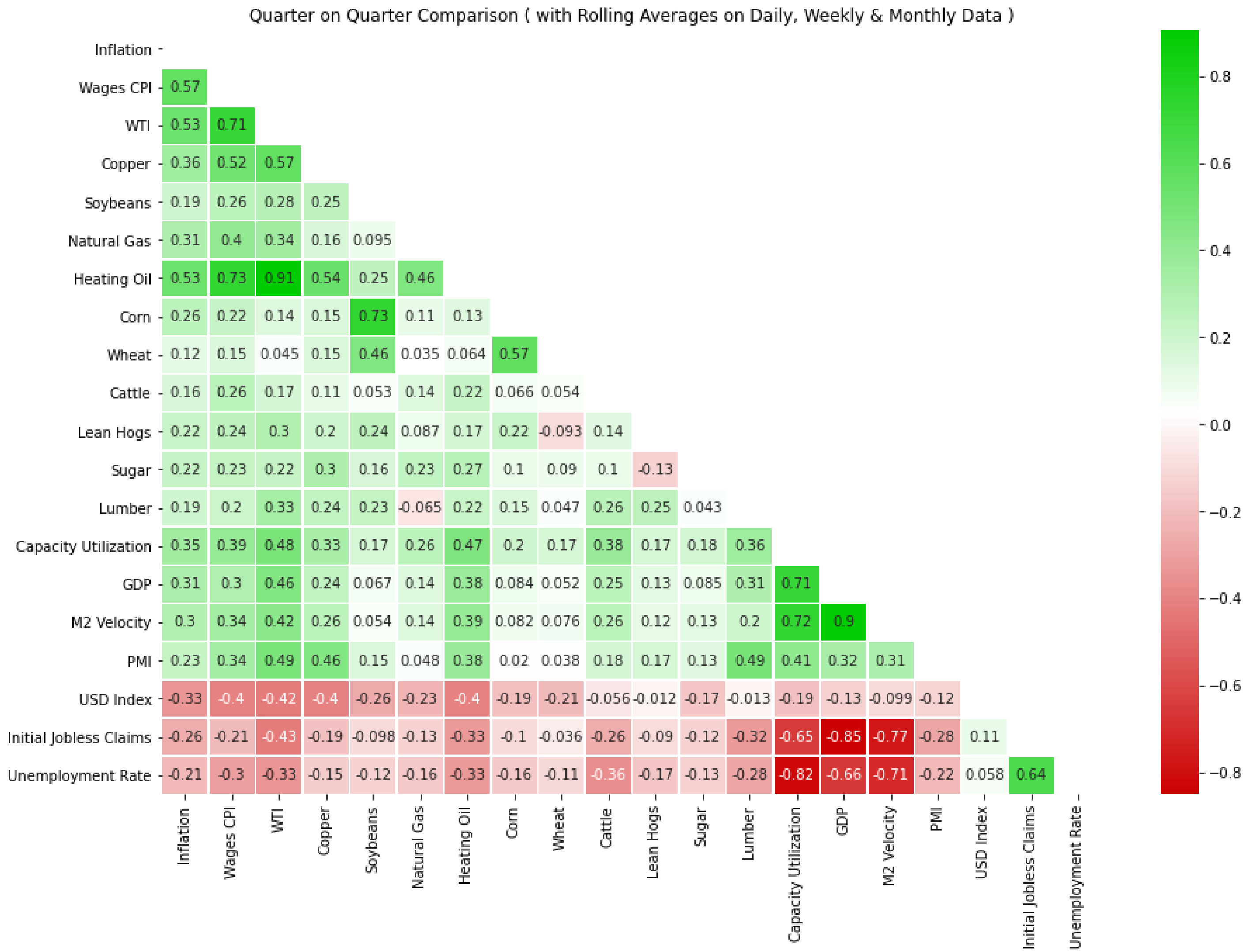
Appendix III (Feature Correlation Heatmap)

#3



Appendix III (Feature Correlation Heatmap)

#4



Appendix IV

```
R² results for nothing scaled below
                    Test 0.254 ( nothing scaled )

R² results for X & y scaled below
SS Train | 0.3966   Test 0.2796
MM Train | 0.0424   Test -0.1085
LG Train | 0.4149   Test -23.8319

R² results for X only scaled below
SS Train | 0.4185   Test 0.254
MM Train | -0.2444  Test -0.0533
LG Train | 0.4142   Test -23.4693


R² results for the LG & SS combination below
SS Train | 0.4067   Test -22.811

R² averages of LG & SS X only scaled below
Av. Train | 0.4164   Test -11.6077
```

~~~~~~

```
MAE results for nothing scaled below
                    Test 0.563 ( nothing scaled )

MAE results for X & y scaled below
SS Train | 0.5376   Test 0.684
MM Train | 0.0811   Test 0.0943
LG Train | 0.5478   Test 1.6306

MAE results for X only scaled below
SS Train | 0.4312   Test 0.563
MM Train | 0.6711   Test 0.6112
LG Train | 0.4381   Test 1.2897


MAE results for the LG & SS combination below
SS Train | 0.4377   Test 1.2751

MAE averages of LG & SS X only scaled below
Av. Train | 0.4346  Test 0.9263
```

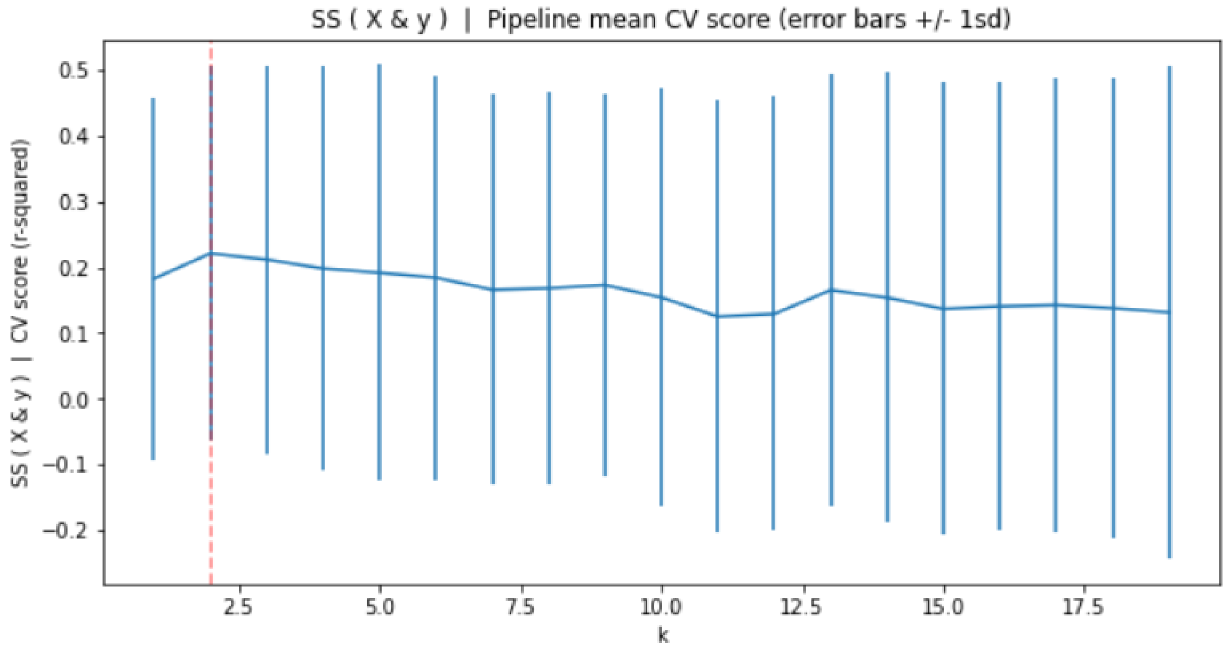
~~~~~~

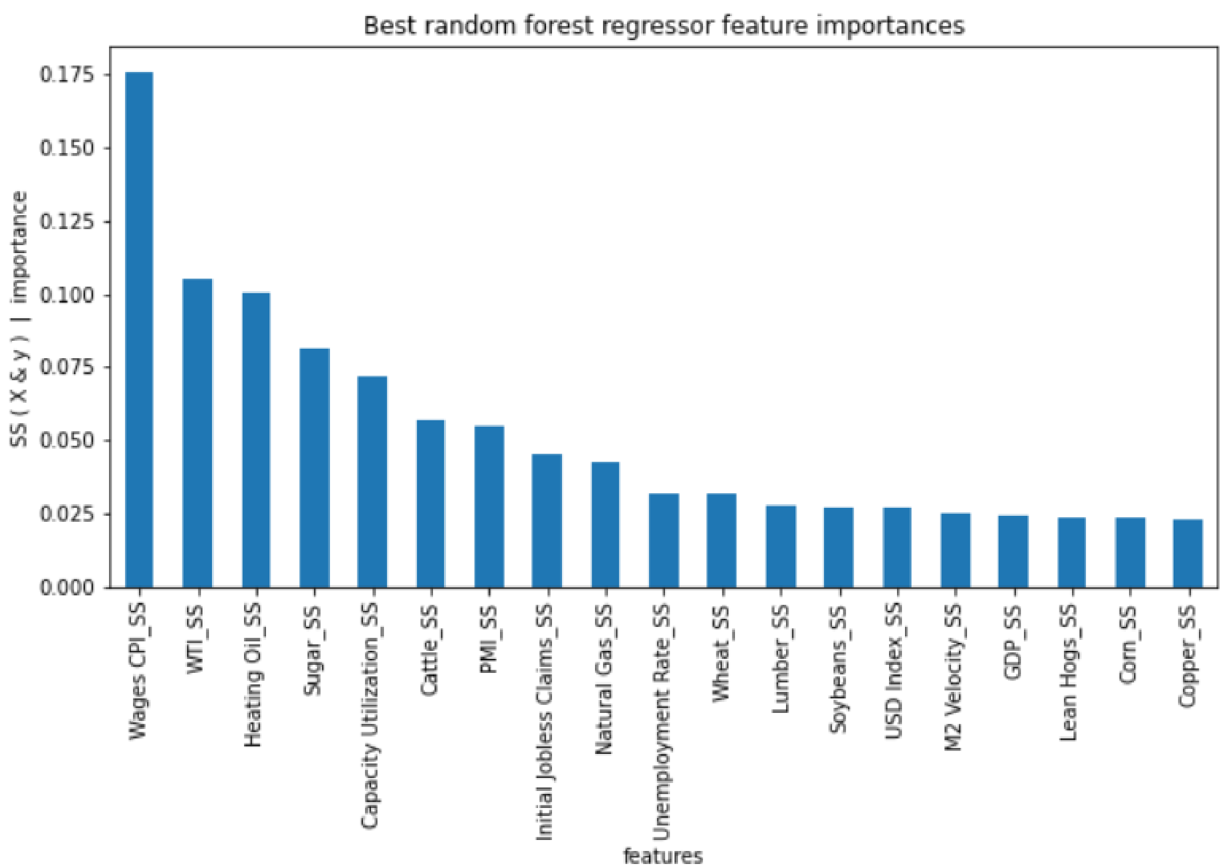
```
MSE results for nothing scaled below
                   Test 0.7556 ( nothing scaled )

MSE results for X & y scaled below
SS Train | 0.6034   Test 0.8602
MM Train | 0.0105   Test 0.0146
LG Train | 0.5851   Test 30.4687

MSE results for X only scaled below
SS Train | 0.3727   Test 0.571
MM Train | 0.7976   Test 0.8061
LG Train | 0.3755   Test 18.7277


MSE results for the LG & SS combination below
SS Train | 0.3803   Test 18.2239

MSE averages of LG & SS X only scaled below
Av. Train | 0.3741  Test 9.6493
```

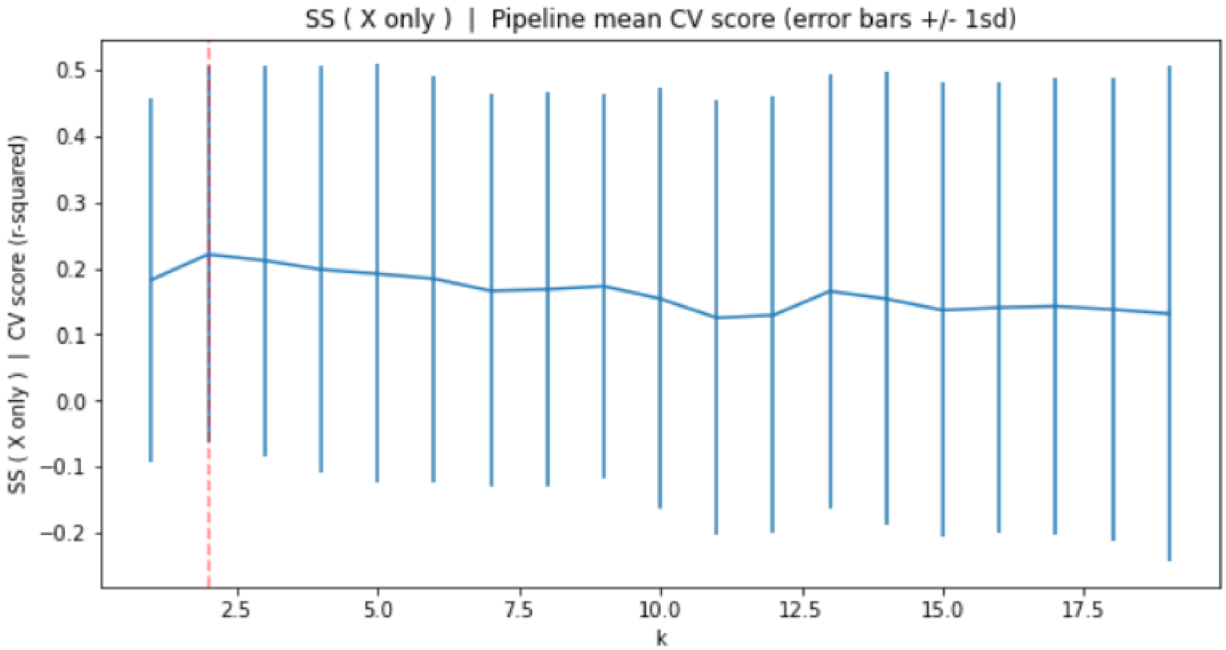
Appendix V

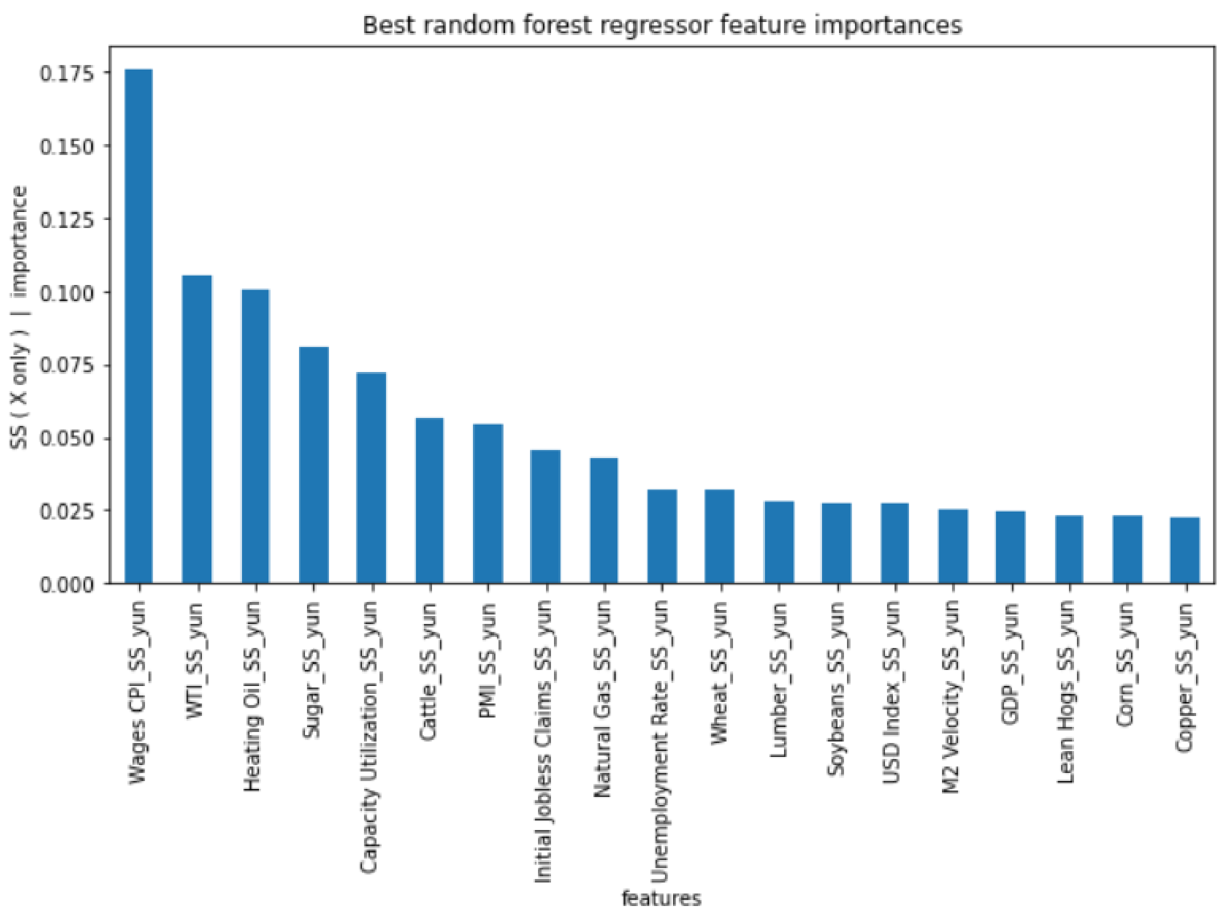
```
R² results for X & y scaled below
SS Train | 0.3966   Test 0.2796
LG Train | 0.4149   Test -23.8319

R² results for X only scaled below
SS Train | 0.4185   Test 0.254
LG Train | 0.4142   Test -23.4693

R² results for the LG & SS combination below
SS Train | 0.4067   Test -22.811
```

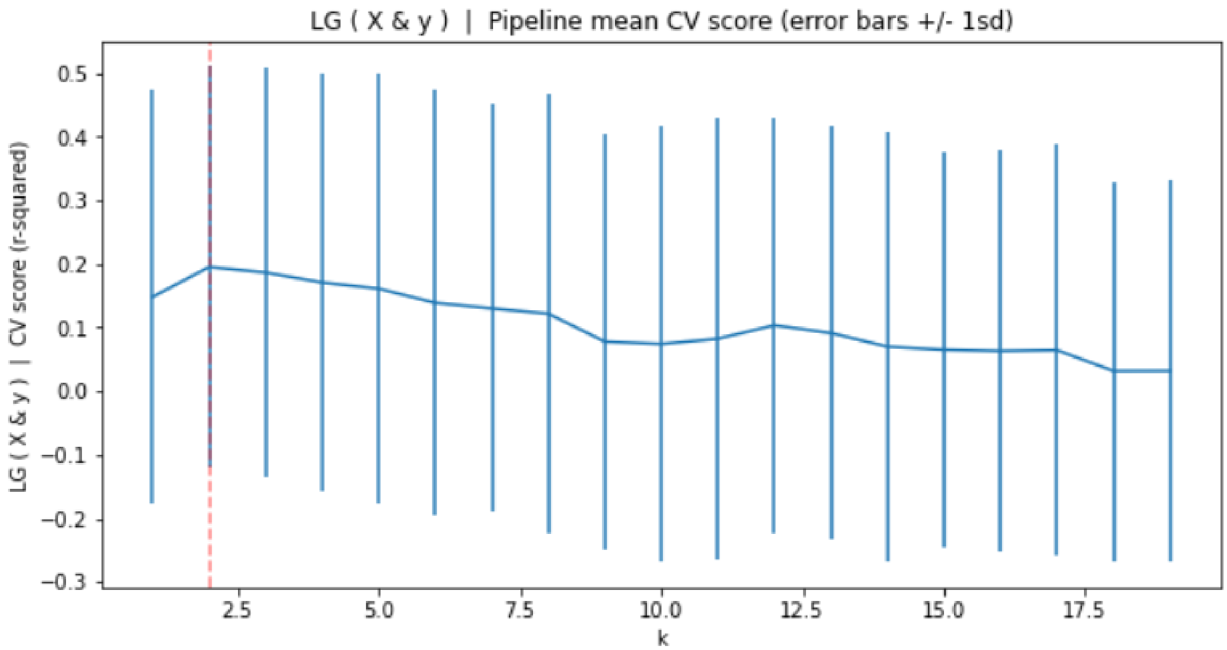
Appendix VI

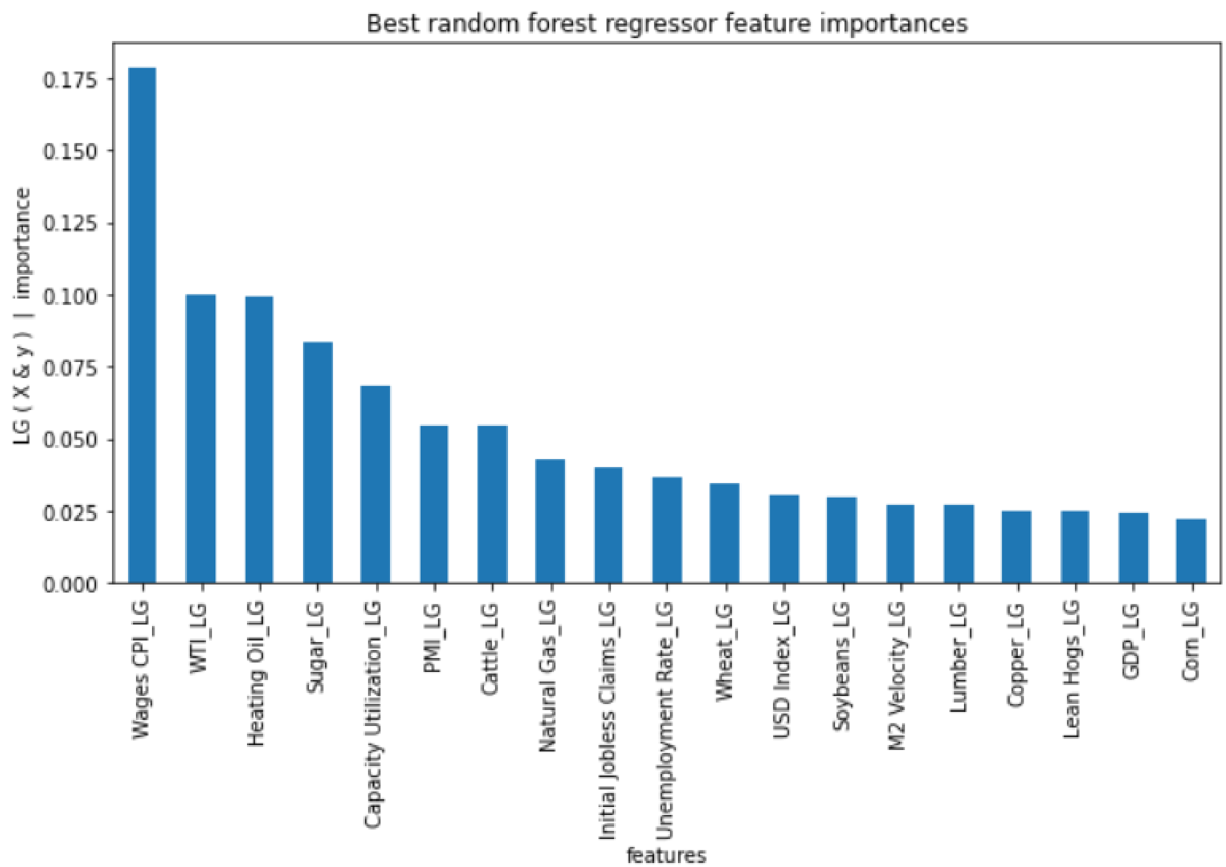

SS (X & y) | Pipeline mean CV score (error bars +/- 1sd)
SS (X & y) | CV score (r-squared)
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
2.5
5.0
7.5
10.0
12.5
15.0
17.5
k

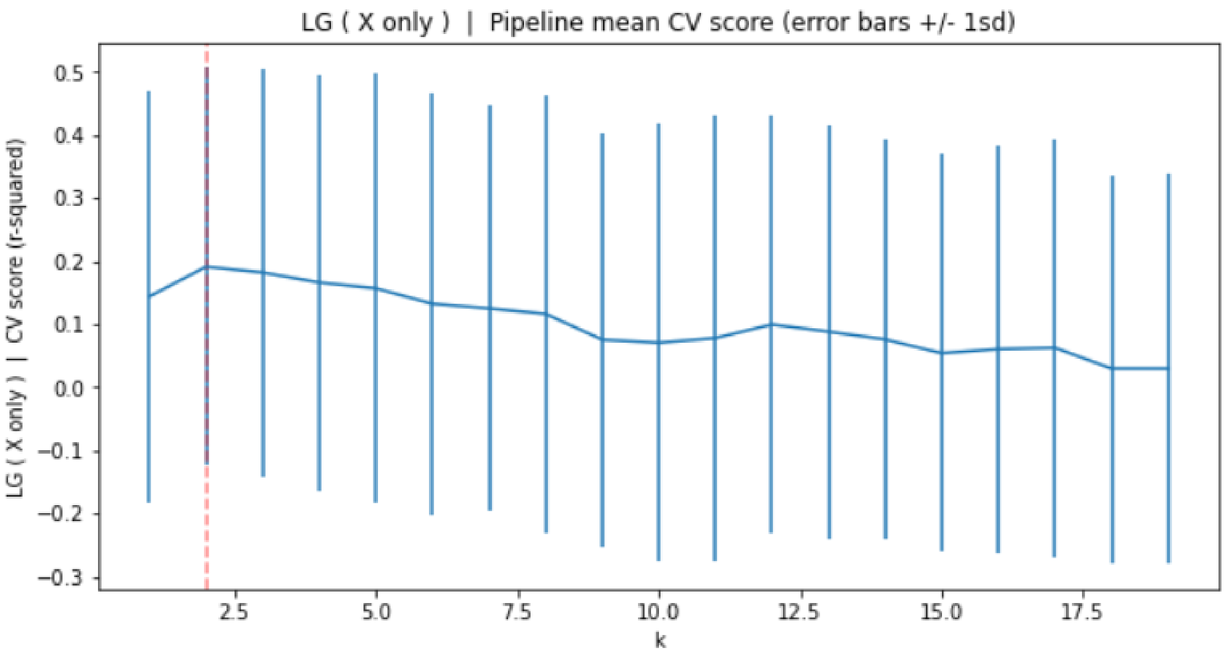

Best random forest regressor feature importances
SS (X & y) | importance
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Wages CPI_SS
WTI_SS
Heating Oil_SS
Sugar_SS
Capacity Utilization_SS
Cattle_SS
PMI_SS
Initial Jobless Claims_SS
Natural Gas_SS
Unemployment Rate_SS
Wheat_SS
Lumber_SS
Soybeans_SS
USD Index_SS
M2 Velocity_SS
GDP_SS
Lean Hogs_SS
Corn_SS
Copper_SS
features


SS (X only) | Pipeline mean CV score (error bars +/- 1sd)
SS (X only) | CV score (r-squared)
k
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
2.5
5.0
7.5
10.0
12.5
15.0
17.5

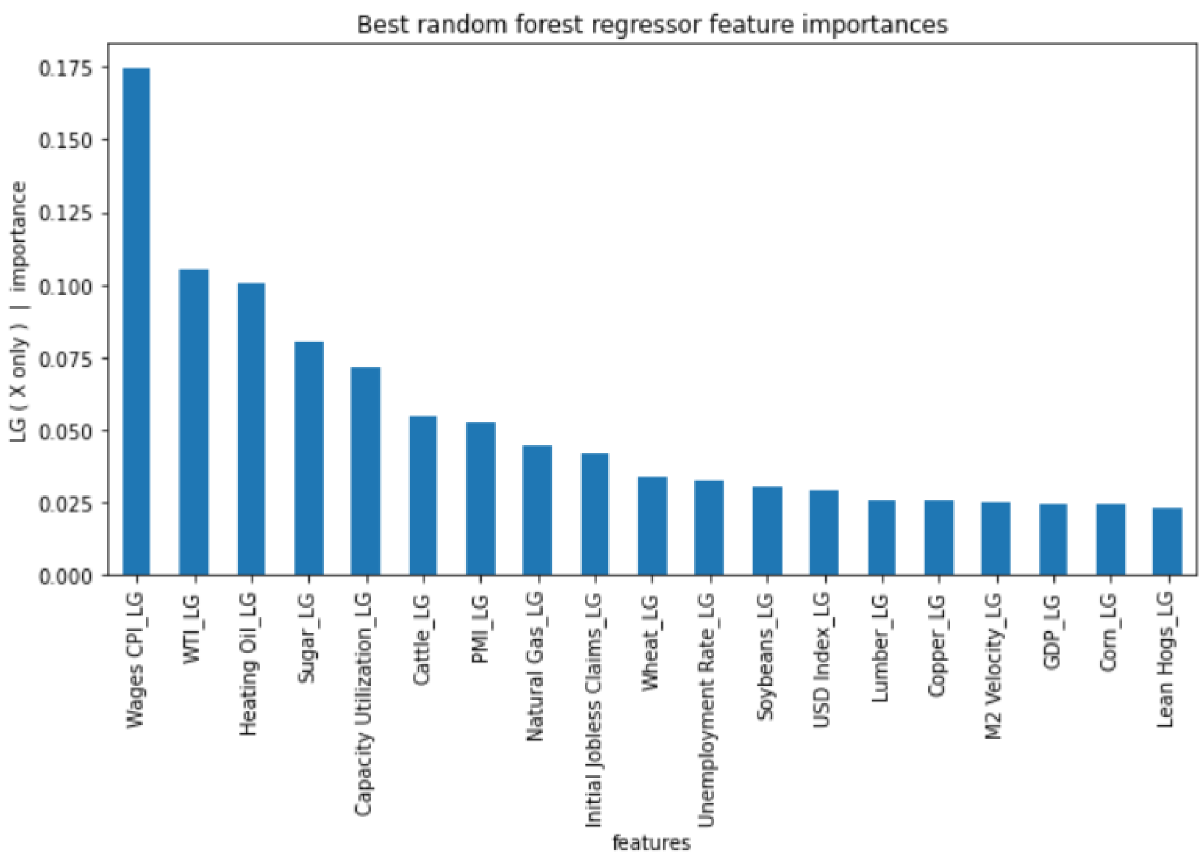

Best random forest regressor feature importances
SS (X only) | importance
features
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Wages CPI_SS_yun
WTI_SS_yun
Heating Oil_SS_yun
Sugar_SS_yun
Capacity Utilization_SS_yun
Cattle_SS_yun
PMI_SS_yun
Initial Jobless Claims_SS_yun
Natural Gas_SS_yun
Unemployment Rate_SS_yun
Wheat_SS_yun
Lumber_SS_yun
Soybeans_SS_yun
USD Index_SS_yun
M2 Velocity_SS_yun
GDP_SS_yun
Lean Hogs_SS_yun
Corn_SS_yun
Copper_SS_yun


LG (X & y) | Pipeline mean CV score (error bars +/- 1sd)
LG (X & y) | CV score (r-squared)
k
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
2.5
5.0
7.5
10.0
12.5
15.0
17.5

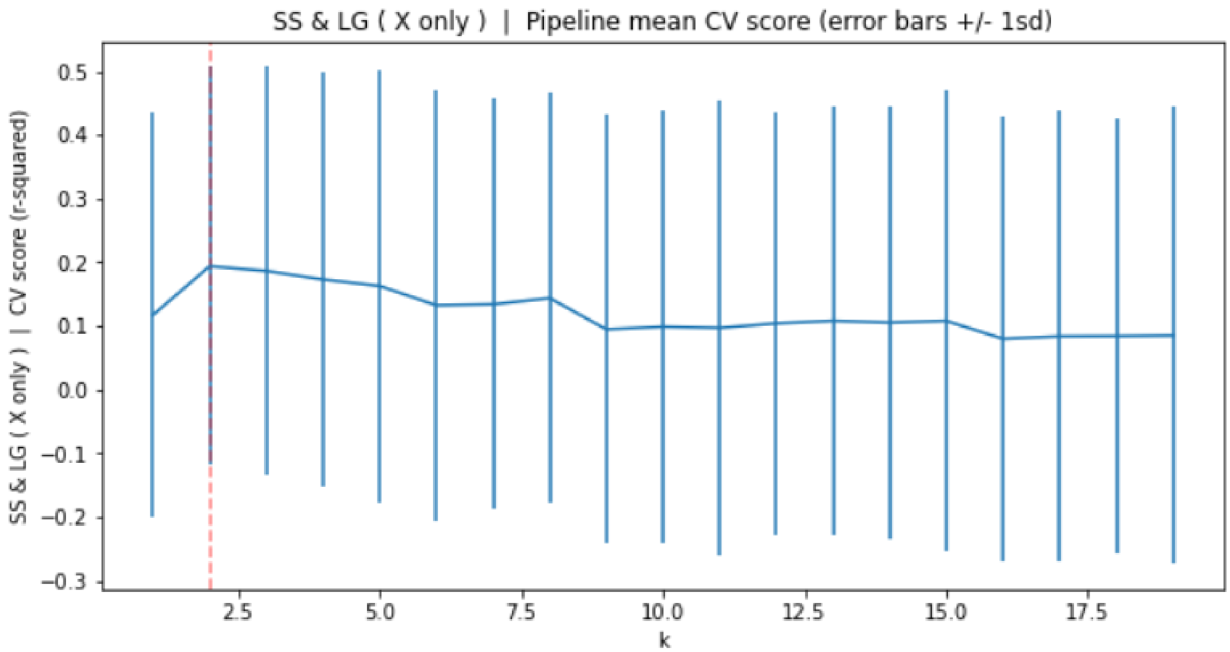

Best random forest regressor feature importances
LG (X & y) | importance
features
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Wages CPI_LG
WTI_LG
Heating Oil_LG
Sugar_LG
Capacity Utilization_LG
PMI_LG
Cattle_LG
Natural Gas_LG
Initial Jobless Claims_LG
Unemployment Rate_LG
Wheat_LG
USD Index_LG
Soybeans_LG
M2 Velocity_LG
Lumber_LG
Copper_LG
Lean Hogs_LG
GDP_LG
Corn_LG

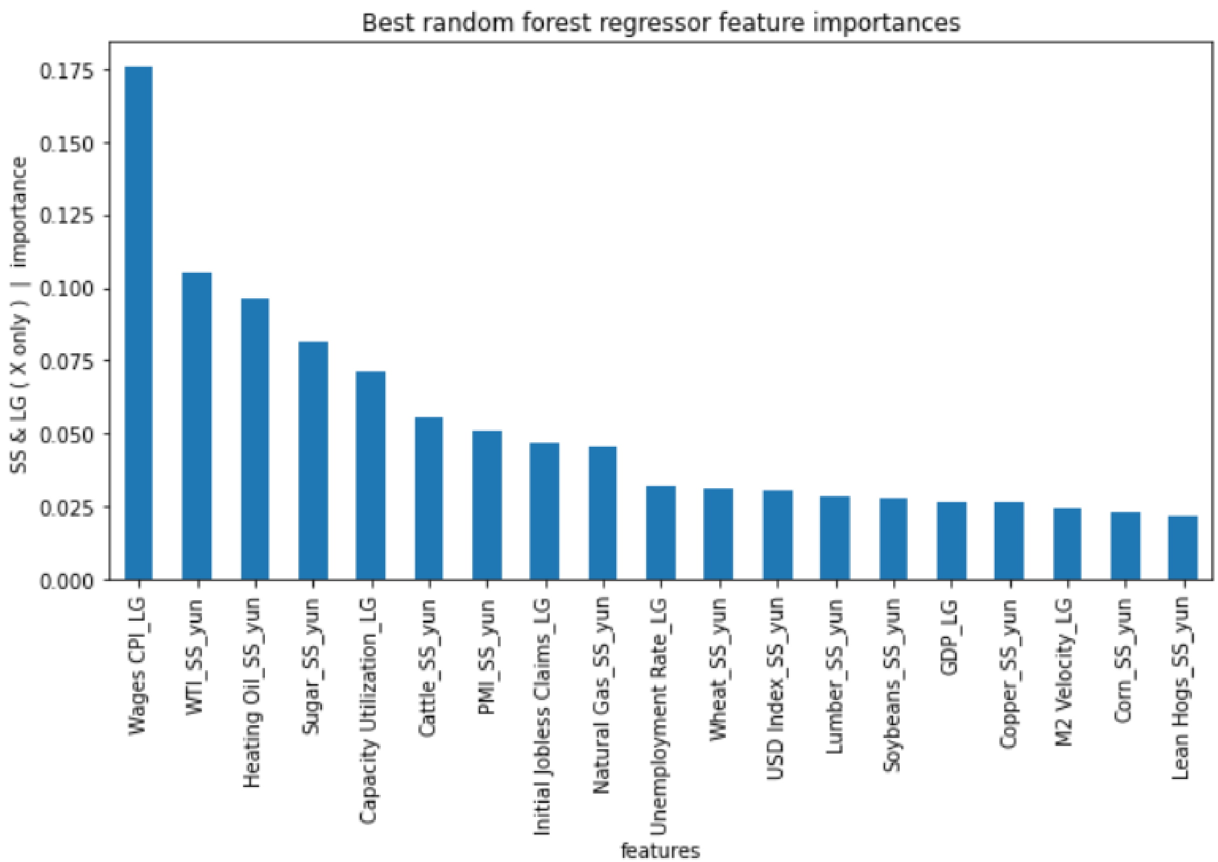


LG (X only) | Pipeline mean CV score (error bars +/- 1sd)
LG (X only) | CV score (r-squared)
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
2.5
5.0
7.5
10.0
12.5
15.0
17.5
k



Best random forest regressor feature importances
LG (X only) | importance
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Wages CPI_LG
WTI_LG
Heating Oil_LG
Sugar_LG
Capacity Utilization_LG
Cattle_LG
PMI_LG
Natural Gas_LG
Initial Jobless Claims_LG
Wheat_LG
Unemployment Rate_LG
Soybeans_LG
USD Index_LG
Lumber_LG
Copper_LG
M2 Velocity_LG
GDP_LG
Corn_LG
Lean Hogs_LG
features


SS & LG (X only) | Pipeline mean CV score (error bars +/- 1sd)
SS & LG (X only) | CV score (r-squared)
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
2.5
5.0
7.5
10.0
12.5
15.0
17.5
k


Best random forest regressor feature importances
SS & LG (X only) | importance
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Wages CPI_LG
WTI_SS_yun
Heating Oil_SS_yun
Sugar_SS_yun
Capacity Utilization_LG
Cattle_SS_yun
PMI_SS_yun
Initial Jobless Claims_LG
Natural Gas_SS_yun
Unemployment Rate_LG
Wheat_SS_yun
USD Index_SS_yun
Lumber_SS_yun
Soybeans_SS_yun
GDP_LG
Copper_SS_yun
M2 Velocity_LG
Corn_SS_yun
Lean Hogs_SS_yun
features