# ECO-395M Final Project: Musab Alquwaee, Daniel Oliner, and Jacob McGill

2024-04-27

## Abstract

We attempted to use data relating to energy production, employment rates, and temperature to predict average energy prices in the US from 2018 to 2023. We conducted an initial analysis looking at the strength of the relationship between energy prices and variables that represented those above factors and found a present but not strong relationship. We then used a linear regression, LASSO regression, and random forest models to attempt to predict the prices in order to determine the best performing model in terms of RMSE. We found that the random forest model had the best performance of the 3.

## Introduction

This project tries to determine if data related to energy generation and economic activity can be used to effectively predict energy prices. Energy prices play a significant role in the world politics and economics. For example, the oil crisis of the 1970s helped spur the stagflation of the decade and more recently the Russian-Ukraine War has significantly impacted Europe by affecting its access to oil and liquid natural gas. Further, energy prices are driven by both supply factors (such as the capacity of power plants and the price of inputs) and demand factors. This project attempts to both identify correlations between energy prices and potential supply side and demand side factors. It also attempts to use those factors to predict the price of energy in cents per kilowatt hour with methods such as LASSO regression and Random Forest. This could provide insight for policy makers and the energy industry on how to predict energy prices and adapt their behavior accordingly.

## Methods

### Data Collection

To construct our data set, we collected data from several sources for the years 2018 to 2023. For the data on monthly energy prices, we used the Energy Information Agency's (EIA) Monthly Electric Power Industry March 2024 report. We also used the EIA's report on monthly energy generation. We also relied on data from the Federal Reserve Economic Data (FRED), including data on the number of Americans employed in the Oil Industry and US Production of Crude Oil. We included data from the National Centers for Environmental Information on average state monthly temperature (in Fahrenheit) and how that temperature deviated from the mean temperature of that state. For state-level monthly employment data, we utilized measures published by the Bureau of Labor Statistics, such as unemployment rates. We also used this data to engineer a new variable capturing monthly state-level labor force participation rates by month. Links for the data sources can be found in the appendix

The variables included in our final dataset are:

- Cents/kWh: The price of energy in cents per kilowatt hour. This is the variable we are trying to predict, the “price” of energy.
- Year, Month, State: The year, month, and state the observation was recorded the observation. We included these variables to account for any seasonal or regional variation in prices. The Month and Year variables were extracted from the “Date” variable using lubridate. The “Date” variable was not included in any of our models.
- Thousand Dollars: Total Revenues collected in the state for electricity generation in thousand dollars. We included these variables to account for revenue received by energy producers
- Megawatthours: Total sales of energy in Megawatt Hours. We included these to account for energy demand.
- Count: Amount of customers serviced in the state. This was also included to account for energy demand.
- Oil_emp(thousands): Amount of individuals employed in the oil industry in the United states. Due to how the data was collected, this did not vary across states, only month and year. We included this variable to account for the quantity of the US energy production.
- U.S. Field Production of Crude Oil: US production of crude oil, measured in thousand of barrels per day. Data for this variable was only available per year, so it was the same by state and month. This was also included to account for the quantity of US energy production.
- Labor Force Participation Rate: Percentage of a state’s eligible working population that was participating in the labor force. As labor force participating is associated with more economic activity and demand, this was included to account for increases in demand from greater economic activity.
- Unemployment Rate: The state’s unemployment rate for that month. This was also included to account for the effects of economic activity on demand.
- Employment-Population Rate: The rate of civilian labor force employed in the state against the total amount of working age population in the state by month. This was also included to account for increases in demand from greater economic activity, such as more people working.
- Gen data: This is data such as gas_gen and hydro_gen that measures the total generation of power by that type in the state measured in megawatthours. The energy types include natural gas, biomass, and solar. We included this data to account for the supply of energy in a given month and state, as well as the influence of the type of energy type.
- Temp and Temp_anomaly: Average temperature in a state for the given month and how much that average temperature deviates from the mean temperature for the state over the time period the data was taken. We included this data to account for variations in demand to weather. This was measured in Fahrenheit.

## Initial Data Analysis

To determine if variables such as energy generation and unemployment rates could predict energy prices, we started by looking at the correlation between energy prices and several variables. We focused on the correlation between the price of energy and unemployment rate, average temperature, and energy generation. We conducted this initial analysis to see if it indicated any predictive relationship between those variables and energy costs. For example, we wanted to see if an increase in unemployment was associated with lower energy prices, if higher energy prices were associated with higher energy costs, or if more energy generation predicted lower prices.

## Model Construction

To predict energy prices, we relied on 3 types of models: a baseline linear regression model, a LASSO regression, and a random forest model. The baseline linear regression model incorporated all available variables in our data set—essentially everything but the kitchen sink. This model would act as the “baseline” to compare the performance of our other other models against it.

We then constructed a LASSO regression model. Since the LASSO model’s regularization technique helps prevent overfitting and enhances the ability to interpret the model, we hoped it could provide a model that would balance interpretability with predictive power. We determined the optimal lambda for the LASSO regression using cross validation.

The third model we used for forecasting was a random forest model. The idea behind this model was to develop a model that would be better at predicting energy prices and could more easily incorporate qualitative features such as states and month than the other 2 models at the cost of interpretability. We developed a random forest model that used all features in our data set.

All models were trained and tested on 80-20 split, with 80% of the data used to train the data and 20% used to test the data. We used Root Mean Square Error (RMSE) to assess the performance of the models.

# Results

We will start by going over the results of our initial data analyis, then the results of our models and their performance.

## Initial Data Analsyis

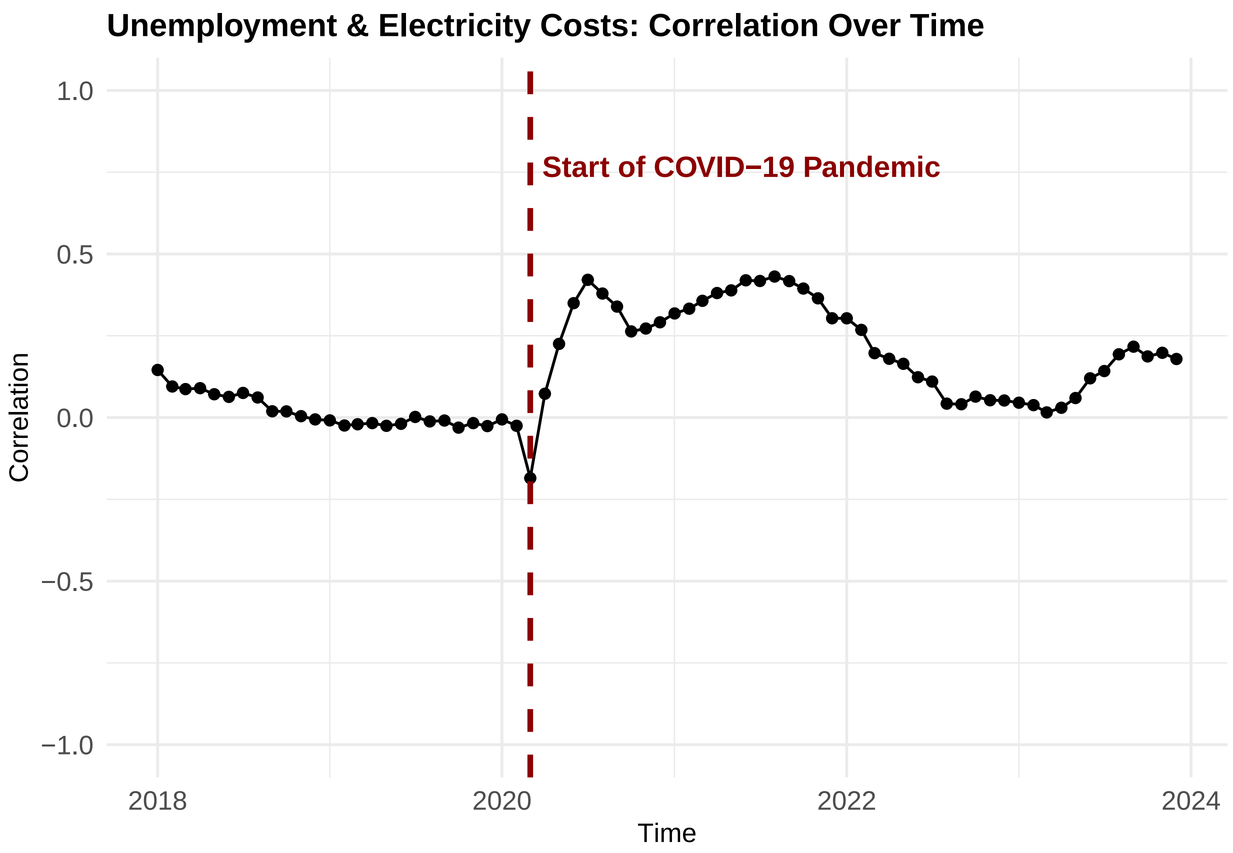
### Unemployment and Energy Prices

To take a preliminary look at the relationship between we conducted a correlation test to determine if a statistically significant correlation between the electricity prices and the unemployment rates exists during this period and the strength of that correlation. Due to the changes to both energy demand and unemployment caused by COVID-19, we excluded 2020 and look at correlation before 2020 and after.

```
##           Estimated Correlation   P-Value
## 2018-2019            0.01659435 0.5855765
## 2021-2023            0.05209927 0.0358430
```

### Table 1

The correlation coefficient of 0.0165 between 2018 and 2019 is a slightly positive relationship but is not statistically significant. The correlation coefficient of 0.052 between 2021 and 2023 is slightly positive and statistically significant. We also looked at how the correlation between energy prices and unemployment changed over time.

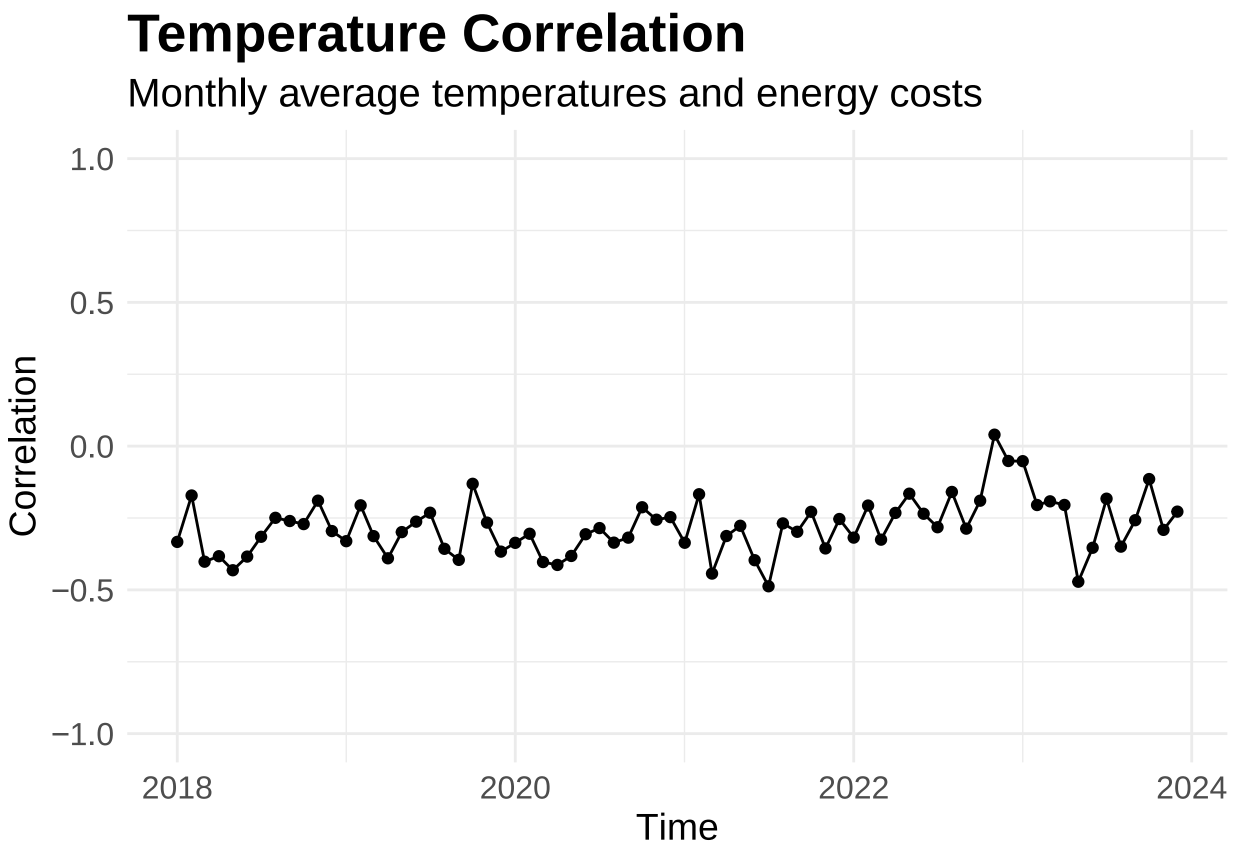


**Graph 1**

This visualization explores the total correlation between unemployment rate and energy costs over time. We see a sharp increase in the correlation between these measures at the onset of the COVID-19 Pandemic in March 2020.

### Temperature and Energy Prices

The below visualization shows the correlation between temperature and energy costs over time, a relationship that is key to understanding how weather influences consumer energy expenditures.



**Graph 2**

**Energy Generation and Energy Prices**

We also looked at the correlation between energy production and energy prices. The first column is the estimated correlation between energy prices in cents per kilowatt hours while the second is the estimate p-value of that correlation.

```
##                         Estimated Correlation       P-Value
## Total Energy Production          -0.159886106  4.963588e-20
## Coal                             -0.406344418 2.597615e-129
## Hydro                            -0.038854249  2.685268e-02
## Gas                              -0.049589533  4.714084e-03
## Other                             0.143846708  1.792864e-16
## Petro                             0.006057674  7.300957e-01
## Solar                             0.234851716  6.425518e-42
## Biomass                           0.319343532  7.423622e-78
## Wind                             -0.163913527  5.519198e-21
## Nuclear                          -0.055622058  1.523199e-03
## Other Gas                        -0.067717932  1.128841e-04
```

**Table 2**

The top line correlation between total energy production and energy prices is negative and statistically significant, although the actual value of the correlation is small, about 0.16. With the exception of petro, all energy generation types have a statistically significant correlation with energy prices, albeit of varying size and direction. About half are negative and half are positive.

## Model Performance

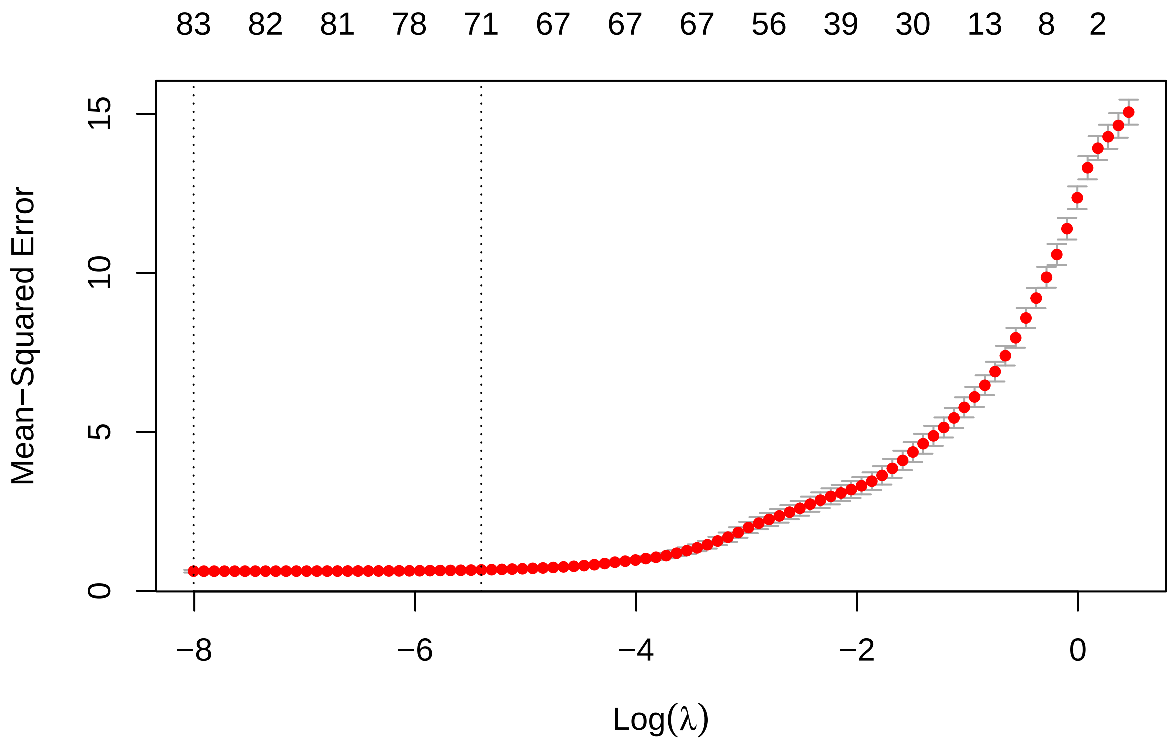
Next we will move onto the results of our models.

### Baseline Model

We will first start with our baseline model. This model is a simple linear regression that we will use its performance as the benchmark for the LASSO and random forest models. It includes all variables in our data set (excluding Date).

### LASSO Regression Results

Next we have the LASSO regression. We again included all variables from our data set, excluding Date. The below visualization of the mean squared error (MSE) from a cross-validation procedure for different values of lambda during LASSO regression shows the identified optimal level of complexity that balances fitting the data well with not being overly complex. The chosen lambda is below the graph. The LASSO regression coefficients are in Table 4 in the appendix.



```
## [1] 0.0003333466
```

**Graph 3**

### Random Forest Results

Finally, we will train the random forest model with all features in the data set, again excluding date. The the inclement mean squared error of each variable is listed in Table 5 in the appendix (Inclement MSE measures by how much a variable improves model performance, so a higher value indicates that a variable is more “important”).

### Model Comparison

Finally, we will move onto to comparing the root mean squared error of the 3 models we tested.

```
##                    RMSE
## Baseline      0.6502221
## LASSO         0.6522432
## Random Forest 0.5989370
```

**Table 3**

The best performing model is the Random Forest, followed by the Baseline, then the LASSO.

# Conclusion

Overall, our results found some correlation between energy prices, temperature, energy generation, and unemployment. For example, the weak correlation coefficients between unemployment and energy price in Table 1 suggests that the predictive relationship between unemployment rate and energy prices is limited. This is further supported by the results in Graph 2, which shows the correlation between energy prices and unemployment shifting over time. There are several potentially overlapping explanations for this insight. First, the pandemic caused a severe economic disruption in the labor market and energy sector, thus the sharp increase in correlation could indicate that these two variables were simultaneously affected by this economic shock. The increase in unemployment caused by the pandemic also led to many people working from home or otherwise altering their daily routine, thus shifting their energy consumption demand patterns.

We also found some relationship between temperature and energy prices, as seen in Graph 2. There, we observed a distinct cyclical pattern in the correlation between these variables, indicating that their relationship is influenced by seasonal factors—likely varying energy demands for heating and cooling throughout the year.

Finally, as can be seen in Table 2, the correlation between energy generation both total and by type varies. Further, they are not very large, although most are statistically significant. These result are counterintuitive; it should be expected that the direction energy generation’s correlation with prices should not vary with type and that energy price and generation have a strong correlation. It may potentially be due to the expansion of some energy types (such as solar) or the cost of generating those energy types: for example if natural gas can be generated quickly but costly, it may have a positive correlation with energy price.

Potential explanations for these overall weaker correlations may be that the relationship between the variable is not very strong (for example unemployment may have a weaker relationship than we initially thought) or more detailed data being required (such as the temperature data being just an average of a state and not capturing the nuances of a state’s environment). There may also be some interactions between these variables that a simple correlation not be effective at identifying.

The results of our models may support that thought, as the random forest model, which allows for more interactions between data by not forcing a simple linear relationship, appears to be the most effective, having the lowest RMSE in our model, as seen in Table 5. However, random forest does not have a substantially higher RMSE. These are also not that relatively large RMSEs. Since price is cents/kWh, these RMSEs are a little more than half a cent/kWh, although considering the scale of energy production, that could be a substantial difference.

The models themselves also have some interesting results. The LASSO model selected a small lambda (see Graph 3), placing small penalty on the model's complexity and keeping most predictors in the model, as seen in the list of coefficients identified by LASSO in the appendix, Table 4. Also, in the random forest model, the variables with the greatest "importance" to the model are coal gen and megawatt hours, as seen in Table 5. The latter makes sense, as it is the total amount of energy generated. Coal is interesting, as it is higher than all other energy generation types and is not that widespread an energy source. Another interesting result of the model is that Temp_Anomaly has the lowest inclemental MSE, since anomalous temperatures could be a strong predictor of higher prices due to unexpected demand. This may be due to the fact that Temp_Anomaly only measures the anomaly as a deviation from the state's average temperature over the entire time period of the data set and not a historical anomaly (such as the difference between the temperature and mean temperature for that month in that state).

There are other limitations to these forecasting models. For example, they require estimates about variables inputted (such as energy generation and unemployment) into the model to make predictions, which may not be easy. It may be difficult for a utility company to predict the amount of wind energy generated in a month to predict prices, which this model requires. Further room for development with this forecasting could include other data that may help predict energy prices, such as amount of new power plants being brought online in the state, capital investment in the state, or the difference between temperature and the historical temperature average for that month. Another approach could be to lag the data; for example, see how energy generation or temperature in a month may predict enery prices into the following month.

# Appendix

## Links for Data

Hyperlinks for each data source are embedded below.

Energy Information Administration: Energy Generation Report, retrieved in March 2024, and Energy Prices Data, retrieved March 2024 from the EIA861M Report, U.S. Crude Oil Production

National Centers for Environmental Information: State Average Temperature Data, collected for each state.

FRED: Oil and Gas Extraction Employment

Bureau of Labor Statistics: Local Area Unemployment Statistic

These individual datasets can be found in the upload_data folder of our github. The code used to clean and join them can be found in the folder cleaning_code. The final data set was created by the R script data_combine.R.

## LASSO Regression coefficients

```
## 85 x 1 sparse Matrix of class "dgCMatrix"
##                                                                           s1
## (Intercept)                                                     3.385605e+01
## (Intercept)                                                     .           
## ...1                                                           -8.400786e-04
## Year2019                                                       -2.573843e-01
## Year2020                                                       -4.711090e-01
## Year2021                                                       -3.651477e-01
## Year2022                                                        4.472905e-02
## Year2023                                                        2.953282e-01
## Month2                                                         1.828162e-01
## Month3                                                         6.658485e-02
## Month4                                                        -1.906611e-02
## Month5                                                        -6.918610e-02
## Month6                                                         1.555032e-01
## Month7                                                         1.255731e-02
## Month8                                                         5.047680e-02
## Month9                                                         8.232832e-02
## Month10                                                       -4.931536e-02
## Month11                                                       -4.964802e-02
## Month12                                                       -8.532012e-02
## StateAL                                                       -8.789310e+00
## StateAR                                                       -1.114982e+01
## StateAZ                                                       -6.651623e+00
## StateCA                                                        7.390563e+00
## StateCO                                                       -5.880121e+00
## StateCT                                                        6.897369e-01
## StateDE                                                       -9.304409e+00
## StateFL                                                        4.851992e-01
## StateGA                                                       -4.877605e+00
## StateID                                                       -1.105204e+01
## StateIL                                                       -2.731908e+00
## StateIN                                                       -7.085368e+00
## StateKS                                                       -7.098950e+00
## StateKY                                                       -1.055582e+01
## StateLA                                                       -1.009248e+01
## StateMA                                                        1.291172e+00
## StateMD                                                       -5.214280e+00
## StateME                                                       -4.871812e+00
## StateMI                                                       -4.352939e+00
## StateMN                                                       -5.201508e+00
## StateMO                                                       -7.749445e+00
## StateMS                                                       -1.064649e+01
## StateMT                                                       -1.059457e+01
## StateND                                                       -1.001027e+01
## StateNE                                                       -8.777608e+00
## StateNH                                                       -4.193340e-01
## StateNJ                                                       -2.818793e+00
## StateNM                                                       -1.126504e+01
## StateNV                                                       -9.546634e+00
## StateNY                                                        9.925174e-01
## StateOH                                                       -4.927953e+00
## StateOK                                                       -9.057275e+00
## StateOR                                                       -9.044812e+00
## StatePA                                                       -3.072455e+00
## StateRI                                                       -9.362347e-01
## StateSC                                                       -8.338766e+00
## StateTN                                                       -7.166822e+00
## StateTX                                                        7.487326e+00
## StateUT                                                       -9.720919e+00
## StateVA                                                       -5.566961e+00
## StateVT                                                       -3.678513e+00
## StateWA                                                       -7.188394e+00
## StateWI                                                       -5.708347e+00
## StateWV                                                       -1.304452e+01
## StateWY                                                       -1.194169e+01
## ‘Thousand Dollars‘                                             4.329634e-06
## Megawatthours                                                 -5.337311e-07
## Count                                                         -9.178498e-07
## ‘oil_emp(thousands)‘                                           2.574383e-02
## ‘U.S. Field Production of Crude Oil Thousand Barrels per Day‘ -3.916083e-05
## ‘Labor Force Participation‘                                   -1.472434e-01
## ‘Employment-Population Rate‘                                  -8.422011e-02
## ‘Unemployment Rate‘                                           -7.503315e-02
## coal_gen                                                       2.146521e-07
## hydro_gen                                                      3.753874e-08
## gas_gen                                                       -1.692317e-08
## other_gen                                                      2.911885e-06
## petro_gen                                                      1.656200e-06
## solar_gen                                                      1.275122e-08
## biomass_gen                                                   -2.217559e-06
## wind_gen                                                      -2.738655e-07
## nuclear_gen                                                   -1.371385e-07
## other_gas_gen                                                 -9.972180e-08
## pump_gen                                                      -2.688934e-06
## Temp                                                           2.046525e-03
## Temp_Anomaly                                                   9.608918e-03
```

**Table 4**

## Random Forest Inclement MSE

```
##                      %IncMSE IncNodePurity
## Year             0.810582230     826.04152
## Month            0.022372517      58.01854
## State            3.565122520    2792.93508
## Thousand_dollars 2.577303162    1777.62351
## Megawatthours    7.660496897    5289.09774
## Count            5.025561319    2911.23179
## oil_emp          0.425350981     440.09247
## US_Oil_Prod      0.155373916     236.52109
## Lab_part         1.399441721    1258.13532
## Employ_rate      0.705736544     615.79347
## UR               0.217795500     261.81470
## coal_gen         7.850453316    8768.34488
## hydro_gen        1.575780012    1268.86103
## gas_gen          0.924006326     830.11618
## other_gen        2.682248393    1715.74578
## petro_gen        0.801081183     914.08259
## solar_gen        1.028020461    1180.57048
## biomass_gen      5.616355591    3024.16019
## wind_gen         3.029062432    1819.78687
## nuclear_gen      1.513222225     692.50994
## other_gas_gen    1.266234912    1063.73795
## pump_gen         0.590158709     444.30428
## Temp             0.105942314     161.40944
## Temp_Anomaly     0.008204396      69.53717
```

**Table 5**