Predictive power of food and gas in atmospheric carbon dioxide concentration

IAQF Competition, February 2021

Abstract—This paper studies the power that various food, gas and oil prices have in determining the level of carbon dioxide parts-per-million (PPM) in the atmosphere. It was found that global energy consumption was the most accurate predictor for carbon concentration among twelve independent variables that were chosen based on their perceived consistent and strong correlation with carbon production and therefore carbon concentration. Lasso and random forest models were run on these variables in order to conclude which inputs had the most predictive power on carbon emissions and eliminate any collinearity between variables chosen. The models were run on both monthly measurements and returns for confirmation purposes where the expectation is that each of the data sets passed in will produce similar results. The outcome from these models was that the best indicator for predicting carbon emissions was global energy, with approximately half of the other variables initially chosen largely being omitted due to collinearity or lack of a strong connection to carbon emissions themselves. Overall, the amount of energy produced by any source–including those that emit less carbon dioxide than regular fossil fuels–is valuable in predicting the amount of emissions and the relative change that can occur over a period of time. The strong relationship between these two factors follow the common consensus of governments' transition towards cleaner energy alternatives. Ultimately, the areas of the world with the most industry and influence need to prioritize stopping the waste of energy use if there is ever an opportunity for the adverse effects of climate change to be curbed.

Index Terms—Linear Regression, Lasso, PPM, Random Forest.

1 Introduction

The publication of Richard Roll's "Orange Juice and Weather" [1] study in The American Economic Review detailed how commodity markets can provide information beyond what is regularly available. Based on these ideas, this paper seeks to determine if other commodity markets could add information or contain more predictive power to what is currently reported with regards to carbon dioxide concentration.

The increasing concentration of carbon dioxide (CO2) in our atmosphere presents a great risk towards society. From well-known environmental impacts such as melting ice caps, rising sea levels, and degradation of arable land to lesser-known impacts across the financial system.

While there's no way to make an immediate change, long term policy adjustments by governments could be vital in maintaining stability within the global market. Extreme weather conditions resulting from climate change will ultimately have a large impact on the world economy through abrupt increases in risk premia across a multitude of assets, as noted in a paper by the Financial Stability Board (FSB) [2]. This could additionally spill into credit, liquidity and counterparty risks. In addition, climate change could lead to an increase in uncertainty in markets, a decrease in asset prices as a result, and cause disruptions to asset classes unrelated such as futures, swaps and other financial instruments.

Aside from the global market as a whole, changes in world climate would have profound effects on world food production, rising-sea level, and protection from radiation. Therefore, it is imperative to make a comprehensive analysis of all possible economic indicators that affect carbon production and to forecast its concentration as accurately as possible. The markets and prices of oil, gas and other commodities are defined by high volatility, in part due to geopolitical reasons and in part due to seasonal supply and demand issues. This means that producers and consumers seek to hedge their prices through various financial instruments such as futures, forward contracts, and swaps. In theory futures prices contain information regarding the expected price and consumption of oil which is an important contributor in carbon production.

2 Preliminary Analysis

To start, commodity markets and consumer level products- in this case West Texas Instruments (WTI) crude oil prices- were analysed and a clear and direct relationship was found, which can be seen below in figure [1].

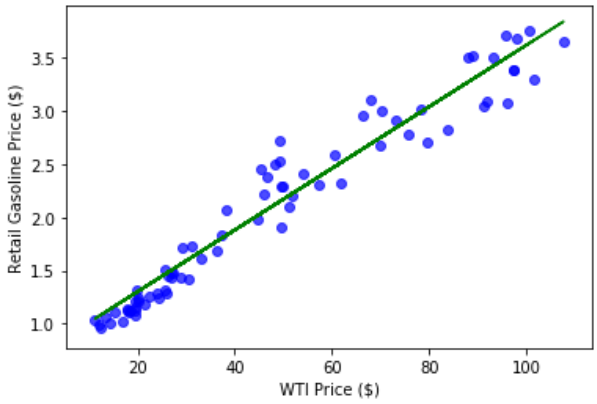


Fig. 1: Relationship between WTI Price and Retail Gasoline Price

The general expectation that increased crude prices reflect increased gasoline prices is not surprising. Retail gasoline is a relatively inelastic product that is required by both individual and commercial consumers, this means that any increase in raw crude prices is passed directly to the consumer by gasoline manufacturers.

However, once WTI crude and CO2 concentration were examined, it can be seen that additional explanatory variables were needed to add colour to the model.

This isn't surprising as CO2 concentration is affected by factors that aren't accounted for. While Roll was able to find a statistically significant relationship between orange juice returns and weather, 97% of orange juice at the time was produced in one area of Florida. On the other hand, oil production is much more evenly spread, from the Gulf countries to Venezuela, Russia and the U.S. This diversity of oil producers introduces factors beyond supply and demand that affect crude oil production and prices, these range from supply chain issues to political factors thus making crude oil a vastly more complex commodity to analyze than orange juice. It can be assumed that there exists a similar relationship between crude oil prices and retail gas prices and crude oil and CO2 concentration due to oil production and consumption being one of the main producers of CO2 in the atmosphere. In actuality the relationship is much more complex.

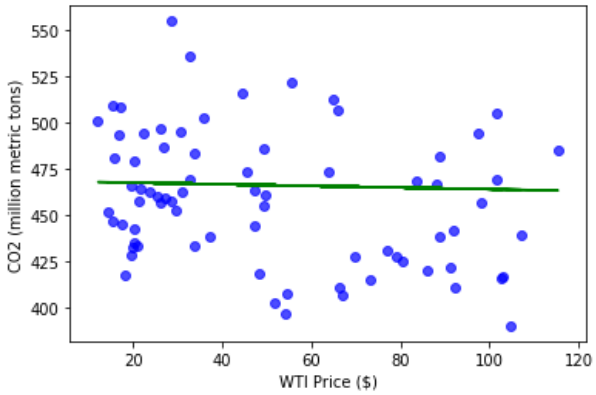


Fig. 2: Relationship between WTI Price and CO2 Concentration

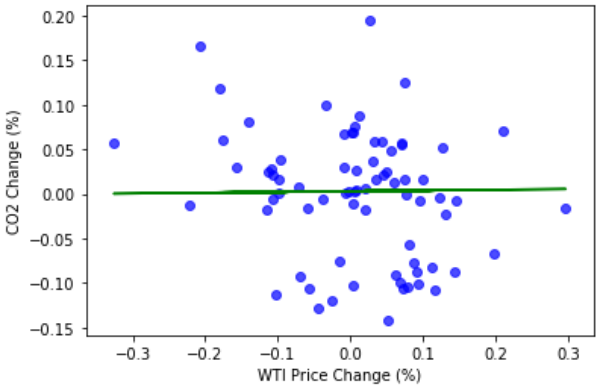


Fig. 3: Relationship between Percent Change WTI Price and Percent Change CO2 Concentration

3 Data Selection

To gauge predictors for carbon emissions, the largest emitting industries were obtained from the U.S. Energy Information Administration's (EIA) monthly energy review [3]. The emissions were measured in millions of metric tons and are further broken down into sectors to be used as a basis for variable selection. The amount that each sector emits is conveyed by Figure [4].

The preeminent variables chosen for the model were global average petroleum stock price (measured in per millions of barrels) and average retail gas price. This was to add the transportation component of the total output. To incorporate electricity

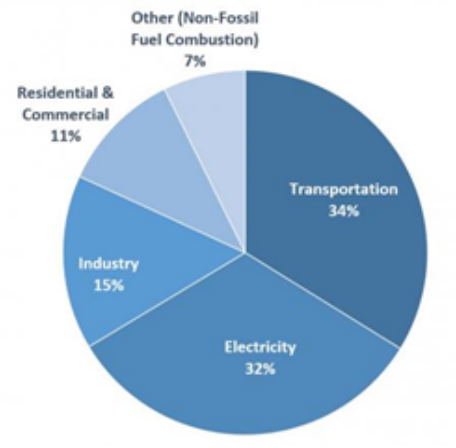


Fig. 4: CO2 Emissions By Sector [4]

into the model, the total output from every known source was included in the model with units of millions of Kilowatts per hour (KWh). Similarly, to account for the emissions from the residential sector, total energy consumption measured in trillions of British Thermal Units (BTU) was selected. This is the total emission from households and businesses alike. In addition, coal was also chosen as a variable due to its prevalence in electricity production as well as in construction. The measure for coal in the model was the average price of coal futures prices as measured by the world bank database [5].

A previous study by Silvennoinen and Thorp explored the correlation between the price of crude oil in relation to a series of agricultural stocks. The paper concluded limited interaction between the energy and food markets with an exception in the extreme scenario where both crude oil and food are priced exceedingly high. Some of the key agricultural prices being tracked in the study included oil, cereal, sugars, meats, and dairy. These are five of the largest components of the Food Price Index (FPI) which is a measure of the monthly change in price for a basket of food commodities [6].

Since agriculture is already associated with carbon emissions (making up for a large part of the non-fossil fuel combustion sector), it's noteworthy to pay attention to oil, cereal, and sugars which are considered biofuel stocks that emit carbon, though at a lower rate than fossil fuels. As the study confirmed a lack of correlation between items within the FPI and crude oil, it's reasonable to believe that the changes in their prices could have predictive power on carbon emissions since production of these commodities directly contributes to the total output in the atmosphere.

4 Statistical Models

The statistical models chosen for the emission estimates are Lasso [7] and Random Forest [8] for their shared purpose of eliminating unnecessary variables and highlighting those with the most predictive power. The models were taken on monthly data spanning 1994-2018 where the data for the model represents global averages or total if specified. Monthly was used due to the unavailability of weekly or daily data.

When running the models, the data was divided in two different ways: randomly and sequentially. In both cases 70% of the data was placed into the training set and the other 30% is used to test the predictive power. When splitting the data sequentially, the first 70% of measurements are used as the training data and the rest of the dates represent the test data for this type of model. These two differences in splits were done to observe the effect that time may have on our predictions and to see if there is a significant difference caused by this.

Both the lasso and random forest regressions were run on two data sets in order to observe whether significant differences between the results of the two groups arose. One model took inputs as the measured monthly values and the other took inputs as monthly percentage changes. Both are significant in determining the variables that can be eliminated and the predictive power of the variables that remain present. Four tests were run in order to incorporate monthly percentage change and value in both time absent and time sensitive models.

4.1 Principle Component Analysis (PCA)

When evaluating the energy based predictors on current emissions values from the past, 28 predictors remain. To find more meaningful insights to these variables PCA [9] was applied to reduce the dimensions of the data when applying the models. As with no lag, there is weaker reasoning to apply PCA, and the importance of each principal component below in figure [5] [10]. From the graph we see the variance explained is comprised of mostly the first four principle components. In figure [6] we see the

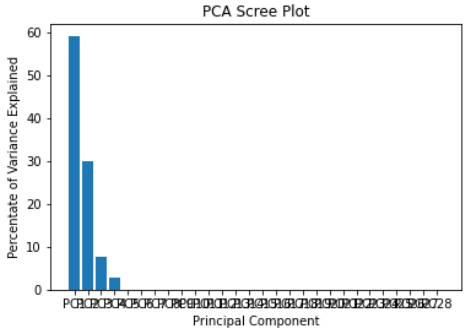


Fig. 5: PCA on data with lag

composition of the variables in the first five principle component.

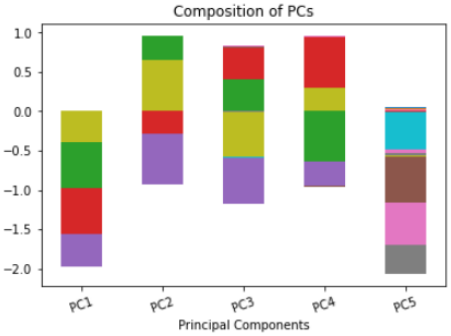


Fig. 6: Composition of the first few Principle Components

4.2 Lasso

The goal of a lasso regression as shown in figure [8] is to minimize the output subtracted by each of the variables respectively. In doing so the model may reduce some of the coefficient Beta values to zero, meaning that those variables had underlying collinearity with some of the other independent factors introduced into the model.

It is also worth considering the large range in magnitude of the values taken within the data set; for retail gas the price varies from below a dollar at certain points while total electricity output exceeds 200,000 throughout the dates taken. Because of this, it's expected that the coefficient for electricity to be in a different order of magnitude to that of the other variables.

For this reason, to interpret the lasso coefficient values, each is multiplied by its mean in order to compare the perceived weight, returning the mean

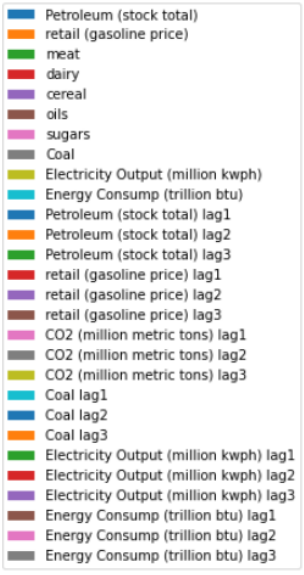


Fig. 7: Legend for Fig. [6]

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

Fig. 8: Lasso Regression Equation

adjusted value. In addition to further standardise the coefficient values, the data was modelled on monthly returns rather than monthly value. The trade-off of using returns rather than monthly values is the potential for higher variation since consecutive monthly values for returns are much more independent than inherent values.

This way the amount that each variable changes in a given month are all comparable to each other and useful for analysis; in addition they're more similar in value to the response since the range of values is much smaller for percentages than for all real values.

The results of the two lasso regressions can be seen below in table [1] and [2] respectively, which correspond to the coefficients of the mean adjusted value and on returns. While the numerical values for each are very different, both models highlight how the variables are to be weighed within our model. In both tables, the highest magnitude variable by a considerable margin is energy, with a coefficient value more than five times bigger than the next biggest coefficient in table [1] and well over 10 times larger than the next highest magnitude value in table [2].

TABLE 1: Random Lasso Coefficient Mean Adjusted Values

R-Squared	0.96154
Percent Error for MSE	0.11114
Petroleum Mean Adjusted Value	-70.01868
Gasoline Mean Adjusted Value	8.79058
Meat Mean Adjusted Value	-18.28847
Dairy Mean Adjusted Value	1.58193
Cereal Mean Adjusted Value	-12.04532
Oils Mean Adjusted Value	0.41817
Sugars Mean Adjusted Value	4.30907
Coal Mean Adjusted Value	-21.98990
Electricity Mean Adjusted Value	0.93930
Energy Consumption Mean Adjusted Value	449.31724

TABLE 2: Random Lasso Coefficient Return Values

R-Squared	0.98295
Percent Error for MSE	1.28321
Petroleum	-0.02468
Gasoline	0.00000
Meat	0.00000
Dairy	0.00146
Cereal	-0.01087
Oils	0.00000
Sugars	-0.00349
Coal	-0.00000
Electricity	0.00000
Energy Consumption	1.00087

TABLE 3: Time Series Lasso Coefficient Mean Adjusted Values

Percent Error for MSE	2.2241e-08
Petroleum Mean Adjusted Value	-0.00225
Gasoline Mean Adjusted Value	0.00089
Meat Mean Adjusted Value	-0.00101
Dairy Mean Adjusted Value	0.00106
Cereal Mean Adjusted Value	-0.00083
Oils Mean Adjusted Value	-0.00041
Sugars Mean Adjusted Value	-0.00055
Coal Mean Adjusted Value	-0.00250
Electricity Mean Adjusted Value	0.00040
Energy Consumption Mean Adjusted Value	0.05889

TABLE 4: Time Series Lasso Coefficient Return Values

MSE	11.35672
Petroleum	-0.01971
Gasoline	0.00000
Meat	-0.00000
Dairy	0.00000
Cereal	-0.00000
Oils	-0.00000
Sugars	-0.00000
Coal	-0.00000
Electricity	0.00333
Energy Consumption	0.99756

The results of the lagged data set without PCA applied with mean adjusted values is shown in table [5]. An improvement in model performance can be seen with the additional variables. In addition, some of the original variables becoming no longer predictive for the response.

When applying PCA we find that the predictive power increases of the variables, however the cost in so many additional variables calls into a problem of over-fitting. We can see the results with the mean adjusted values shown in table [6].

Gasoline, meat, oil, coal, and electricity were among the variables from the monthly returns that the lasso regression omitted due to known correlations with other independent variables. Previous empirical studies have highlighted an asymmetric connection between gasoline and petroleum markets that contains lag which is likely the reason it was deemed redundant by the model. Coal and electricity were both dismissed as necessary variables most likely because of their correlation with each other

TABLE 5: Lagged Lasso Coefficient Mean Adjusted Values (values with 0 not shown)

MSE	0.02397
Petroleum (stock total)	-4.301245
retail (gasoline price)	0.104217
meat	-3.996711
dairy	0.854102
cereal	-4.090819
oils	2.311414
Electricity Output (million kwph)	15.036752
Energy Consump (trillion btu)	462.487253
Petroleum (stock total) lag2	-17.654237
CO2 (million metric tons) lag1	316.625275
CO2 (million metric tons) lag3	1.690026
Coal lag2	-7.877300
Electricity Output (million kwph) lag3	-4.306824
Energy Consump (trillion btu) lag1	-330.183279
Energy Consump (trillion btu) lag2	-1.036994

TABLE 6: PCA Lasso Coefficient Mean Adjusted Values (values with 0 not shown)

MSE	6.40786e-07
PC1	2.363994e-19
PC2	4.158332e-17
PC3	1.570395e-17
PC4	-5.200687e-17
PC5	-1.170399e-16
PC6	-7.129994e-18
PC7	-7.772673e-17
PC8	3.968117e-17
PC9	9.670981e-17
PC10	6.439821e-18
PC11	5.077815e-18
PC12	-4.319731e-18
PC13	3.585919e-17
PC14	3.459121e-17
PC15	4.559467e-17
PC16	1.328244e-17
PC19	1.905870e-17
PC20	-2.768158e-17
PC21	-2.559354e-17
PC22	3.330592e-17
PC23	3.206420e-18
PC24	1.946149e-17
PC25	-2.957770e-18
PC27	-2.399843e-18
PC28	-1.165137e-18

and their underlying relationship with energy output which also includes processes that create electricity using coal. The two food commodities that the algorithm removed were vegetable oil price and meat price. This is expected as these have little impact on carbon output compared to the other foods included.

4.3 Random Forest

Random forest algorithms iterate through multiple decision trees that stem from the training data. For each tree, the random forest method looks for the best independent variable among a random subset in order to determine how the decision tree is to split thereon out. The specific independent variable and observation chosen by the tree are the observation that occurs the most within the subset. Each time a tree is created the random subset is observed to see if there are any values present that can split the data further which will produce further trees and more accurately represent the desired results.

Similar to the analysis using the lasso model, more extreme results are expected when observing the forest for the monthly values rather than the returns. Since the subset of variables and the observations for each standalone tree are random, the way to determine the leading variables in prediction power is through their relative "importance" to each other.

The value of importance for an independent variable is the change in performance if the variable were removed. The higher the value, the higher the importance, the larger the change in performance. Any values of importance value zero are said to be insignificant in their effect on performance of the predictive model. The expectations of the random

forest are to reiterate the results of lasso and grant importance to the variables through which the lasso model granted higher magnitude coefficient values. The results of the analysis are displayed in tables [7]-[12] with their plots located in the appendix; each table indicates the error, accuracy, and relative importance of each variable.

The random forest model produced results using 1000 trees to ensure that there would be enough data for the model to determine the relative importance of each of the chosen variables. Each monthly value result and monthly return result mirror the output provided by the lasso regression model in terms of the weights that are provided to each of the variables. Once again, energy consumption's importance was magnitudes larger than for the other variables; it's likely that without this variable the high values of accuracy wouldn't have been attainable. The accuracy of the returns model is close to being 100% with close to zero error when it comes to the predicted value of CO2 emissions. The increased accuracy of this model in addition to the added randomness that stems from random subset selection further provides evidence of the overpowering predictor of CO2 emissions being energy output. Even with extremely different magnitude values present within their respective sets of data, both the monthly values and returns confirmed the case for energy's necessity to be present within our model.

TABLE 7: Random Forest Diagram: Random Monthly Value

MSE: 123.77514	
Energy with importance	0.96
Oils with importance	0.01
Coal with importance	0.01
Electricity with importance	0.01
Petroleum with importance	0.0
Gasoline with importance	0.0
Meat with importance	0.0
Dairy with importance	0.0
Cereal with importance	0.0
Sugars with importance	0.0

It's interesting to note the discrepancies that take place within the two random forest models that don't occur in lasso. The monthly mean adjusted estimates for lasso and random forest had similar outputs in terms of the values given to coefficients and importance relative to other variables. The lasso ran on monthly returns relayed the same result that energy was the leading predictor, however it left in around half of the other variables as justifiably having predicting power regarding the dependent variable. The random forest for the returns on the hand gave energy an importance score of .96 which is very large and also was repeatable with further random samples. There's also a notable difference in the MSE values for between the monthly value group

TABLE 8: Random Forest Diagram: Random Monthly Returns

MSE: 1.39120	
Energy Consumption	0.68
Petroleum Stock	0.17
Meat	0.04
Sugars	0.04
Coal	0.03
Electricity Output	0.02
Retail Gasoline	0.01
Dairy	0.01
Cereal	0.01
Oils	0.01

TABLE 9: Random Forest Diagram: Time Series Monthly Value

MSE: 1162.62512	
Energy with importance	0.98912
Electricity with importance	0.00229
Meat with importance	0.0014
Coal with importance	0.00128
Cereal with importance	0.00110
Sugars with importance	0.00108
Petroleum with importance	0.00106
Oils with importance	0.00101
Dairy with importance	0.000987
Gasoline with importance	0.00064

TABLE 10: Random Forest Diagram: Time Series Monthly Returns

MSE: 1.75659	
Energy with importance	0.14988
Dairy with importance	0.0.11293
Meat with importance	0.10677
Electricity with importance	0.10617
Oils with importance	0.10370
Gasoline with importance	0.09655
Coal with importance	0.08326
Cereal with importance	0.08328
Sugars with importance	0.07652
Petroleum with importance	0.07562

TABLE 11: Lagged Random Forest Diagram: Monthly Values (values with 0 not shown)

MSE: 0.03868	
Energy Consump (trillion btu) with importance	0.69
Petroleum (stock total) with importance	0.12
CO2 (million metric tons) lag1 with importance	0.05
meat with importance	0.03
Petroleum (stock total) lag1 with importance	0.02
sugars with importance	0.01
Electricity Output (million kwph) with importance	0.01
Petroleum (stock total) lag2 with importance	0.01
Petroleum (stock total) lag3 with importance	0.01
CO2 (million metric tons) lag3 with importance	0.01
Coal lag2 with importance	0.01
Coal lag3 with importance	0.01
Energy Consump (trillion btu) lag3 with importance	0.01

and the time series data set; the monthly values produced a much larger MSE which could be sign pointing to the percent change being the better of the two inputs for prediction power. These differences are consistent when the method of picking training and testing sets is changed as well so it's reasonable that the percent change models (which deem most of the independent variables as redundant and unnecessary) may be more valuable and accurate in their predictions

TABLE 12: PCA Random Forest Diagram: Monthly Values (values with 0 not shown)

MSE: 0.00034	
PC7 with importance	0.34
PC6 with importance	0.23
PC5 with importance	0.11
PC2 with importance	0.06
PC8 with importance	0.05
PC22 with importance	0.03
PC1 with importance	0.02
PC3 with importance	0.02
PC4 with importance	0.01
PC9 with importance	0.01
PC10 with importance	0.01
PC11 with importance	0.01
PC12 with importance	0.01
PC13 with importance	0.01
PC14 with importance	0.01
PC15 with importance	0.01
PC16 with importance	0.01
PC17 with importance	0.01
PC18 with importance	0.01
PC19 with importance	0.01
PC21 with importance	0.01
PC23 with importance	0.01
PC24 with importance	0.01
PC25 with importance	0.01
PC26 with importance	0.01
PC27 with importance	0.01

5 Results and Discussion

Each of the models chosen for our analysis performed as predicted, although the results weren't aligned with the diversification expected prior to running each of the tests. Through the research done prior to running the models, the selection of data was done so as to cover every possible source of carbon emission to reduce any possible noise that may come from a source that was undesirable. In the case of the variables introduced to predict CO2 emissions, for the most part commodities located within the sectors that produce the most carbon emissions seem to not

have predictive power in the amount of CO2 even though they are the direct cause of it.

Testing the models using both actual values which contained a large range of magnitudes and using percent values were useful in providing consistent results in the form of variable weights. The training and testing sets that were chosen based on time were stricter because the lasso and random forest deemed many of the inputs unnecessary which was less readily apparent when the sets were split randomly. However, the general conclusion that came from each test was uniform in the direction that it points in; predominantly that there was a large amount of collinearity between the initial set of independent variables passed in and removing them would not make much of a difference.

The fact that the models uniformly produced energy output as the best predictor is predictable due to its inherent relationship with carbon emissions and because energy consumption has no boundaries in terms of what sectors contribute to energy output. In addition, energy is produced in all industries around the world, potentially explaining why the other variables were reduced in comparison and why the overlap between energy and many of the independent variables is not surprising. This was likely a contributing factor in dismissing the variables thrown out by the models.

6 Conclusion

The selection of lasso and random forest models respectively were effective in eliminating variables that seemed to be unnecessary in predicting total carbon emission output, however the predictors and their relative performances came as a surprise. The independent variable that had overwhelming predictive power compared to the other selected variables was energy output, even more so when returns were observed rather than actual values. While the strong connection between energy production and carbon emissions is not hard to imagine, it was surprising the magnitude of correlation.

In looking towards the future, the clear solution is finding alternatives to our current methods of energy consumption that don't cause adverse effects to the environment. Since this is a metric that has global implications, there needs to be a global cooperative initiative to slow down the pace that climate change is occurring at, at which the Paris Climate Accord attempts to solve. The requirement is for every global entity to transition to a more environmentally friendly source of energy in order to insure safer and stable living conditions for generations to come.

References

[1] R. Roll, "*Orange Juice and Weather*," *The American Economic Review*, vol. 74, pp. 861–880, 1984.

[2] F. S. Board, "The implications of climate change for financial stability." https://www.fsb.org/wp-content/uploads/P231120.pdf.

[3] E. I. Administration, "Monthly energy review." https://www.eia.gov/totalenergy/data/monthly/.

[4] E. P. Agency, "Inventory of u.s. greenhouse gas emissions and sinks." https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks.

[5] W. Bank, "World bank database." https://databank.worldbank.org/home.aspx.

[6] A. Silvennoinen and S. Thorp, "*Crude Oil and Agricultural Futures: An Analysis of Correlation Dynamics*," *The Journal of Futures Markets*, vol. 36, pp. 522–544, 2016.

[7] Smith, "Ridge regression and the lasso." http://www.science.smith.edu/ jcrouser/SDS293/labs/lab10-py.html.

[8] K. W, "Random forest in python." https://towardsdatascience.com/random-forest-in-python-24d0893d51c0.

[9] S. learn, "Scikit decomposition." https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.

[10] Pythontic, "Inventory of u.s. greenhouse gas emissions and sinks." https://pythontic.com/pandas/dataframe-plotting/bar.

Appendix 1

```python
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split


f=pd.read_excel("using_total_electricity.xlsx")
x=f["WTI Price"]
y=f["retail (gasoline price)"]
z=f["CO2 (million metric tons)"]
g=f.dropna()
a=g["WTI Price Change %"]
b=g["CO2 Change %"]

#WTI to Retail
x_train, x_test, y_train, y_test = train_test_split(x, y)
LR = LinearRegression()
LR.fit(x_train.values.reshape(-1,1), y_train.values)
prediction = LR.predict(x_test.values.reshape(-1,1))
plt.plot(x_test.values.reshape(-1,1), prediction, label ='Linear Regression', color = 'g')
plt.scatter(x_test, y_test, label = 'test', color = 'b', alpha = .7)
plt.xlabel("WTI Price ($)")
plt.ylabel("Retail Gasoline Price ($)")
plt.show()

#WTI to CO2
x_train, x_test, z_train, z_test = train_test_split(x, z)
LR = LinearRegression()
LR.fit(x_train.values.reshape(-1,1), z_train.values)
prediction = LR.predict(x_test.values.reshape(-1,1))
plt.plot(x_test.values.reshape(-1,1), prediction, label ='Linear Regression', color = 'g')
plt.scatter(x_test, z_test, label = 'test', color = 'b', alpha = .7)
plt.xlabel("WTI Price ($)")
plt.ylabel("CO2 (million metric tons)")
plt.show()

#WTI % to CO2 %
a_train, a_test, b_train, b_test = train_test_split(a, b)
LR = LinearRegression()
LR.fit(a_train.values.reshape(-1,1), b_train.values)
prediction = LR.predict(a_test.values.reshape(-1,1))
plt.plot(a_test.values.reshape(-1,1), prediction, label ='Linear Regression', color = 'g')
plt.scatter(a_test, b_test, label = 'test', color = 'b', alpha = .7)
plt.xlabel("WTI Price Change (%)")
plt.ylabel("CO2 Change (%)")
plt.show()
```

Preliminary analysis code

Appendix 2

```
from sklearn.decomposition import PCA
from sklearn import preprocessing
import matplotlib.pyplot as plt
pca_cols = ['PC'+str(i) for i in range(1,29)]
pca = PCA(n_components = 28)
principalComponents = pca.fit_transform(x1)
principalDataframe = pd.DataFrame(data = principalComponents, columns = pca_cols)
```

```
percent_variance = np.round(pca.explained_variance_ratio_* 100, decimals =2)
plt.bar(x= range(1,29), height=percent_variance, tick_label=pca_cols)
plt.ylabel('Percentate of Variance Explained')
plt.xlabel('Principal Component')
plt.title('PCA Scree Plot')
plt.savefig('PCA.png')
plt.show()
```

```
columns = df.columns
for col in columns:
    if col not in ['fpi', 'meat', 'dairy', 'cereal', 'oils', 'sugars']:
    #df[col + ' pct Change'] = df[col].pct_change()
        df[col + ' lag1'] = df[col].shift(1)
        df[col + ' lag2'] = df[col].shift(2)
        df[col + ' lag3'] = df[col].shift(3)
columnList = list(df.columns)
```

```
pca_df.iloc[:5].plot.bar(stacked=True,rot=20, title="Composition of PCs")
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)
plt.savefig('PCA_makeup.png')
plt.show(block=True)
```

PCA code

Appendix 3

```
get_ipython().magic('matplotlib inline')
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import scale
from sklearn.model_selection import train_test_split
from sklearn.linear_model import Ridge, RidgeCV, Lasso, LassoCV
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import RepeatedKFold
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score
import pydot

f=pd.read_excel("using_total_electricity.xlsx")
f2=pd.read_excel("returns.xlsx")
y=f["CO2 (million metric tons)"]
y_ret=f2["CO2"]
x=f.drop("CO2 (million metric tons)",axis=1)
x1=x.drop("fpi",axis=1)
x_ret=f2.drop("CO2",axis=1)
#values
x_train=x1[:210]
y_train=y[:210]
x_test=x1[210:]
y_test=y[210:]
#returns
x_train=x_ret[:210]
y_train=y_ret[:210]
x_test=x_ret[210:]
y_test=y_ret[210:]
#needed for random testing and training
#X_train, X_test , y_train, y_test = train_test_split(x1, y, test_size=0.7, random_state=17)
lasso = Lasso(max_iter = 100000, normalize = True)
coefs = []
alphas = 10**np.linspace(10,-5,100)*0.5
for a in alphas:
    lasso.set_params(alpha=a)
    lasso.fit(scale(x_train), y_train)
    coefs.append(lasso.coef_)

ax = plt.gca()
ax.plot(alphas*2, coefs)
ax.set_xscale('log')
plt.axis('tight')
plt.xlabel('alpha')
plt.ylabel('weights')
```

Lasso code

```
lassocv = LassoCV(alphas = None, cv = 10, max_iter = 100000, normalize = True)
lassocv.fit(x_train, y_train)
lasso.set_params(alpha=lassocv.alpha_)
lasso.fit(x_train, y_train)
MSE=mean_squared_error(y_test, lasso.predict(x_test))
for value in y_test:
    y+=value
y/=len(y_test)

print("Percent error for MSE: "  + str(MSE/np.mean(y_test)))
p=pd.Series(lasso.coef_, index=x1.columns)
mean_list=[339.7078567,2.22154,81.75566667,85.08333333,86.10566667,88.40766667,81.95133333,142.9,324164.5492,8068.911737]
zipped=list(zip(mean_list,p))
new=[]
for m,v in zipped:
    new.append(m*v)

pd.Series(new,index=x1.columns)

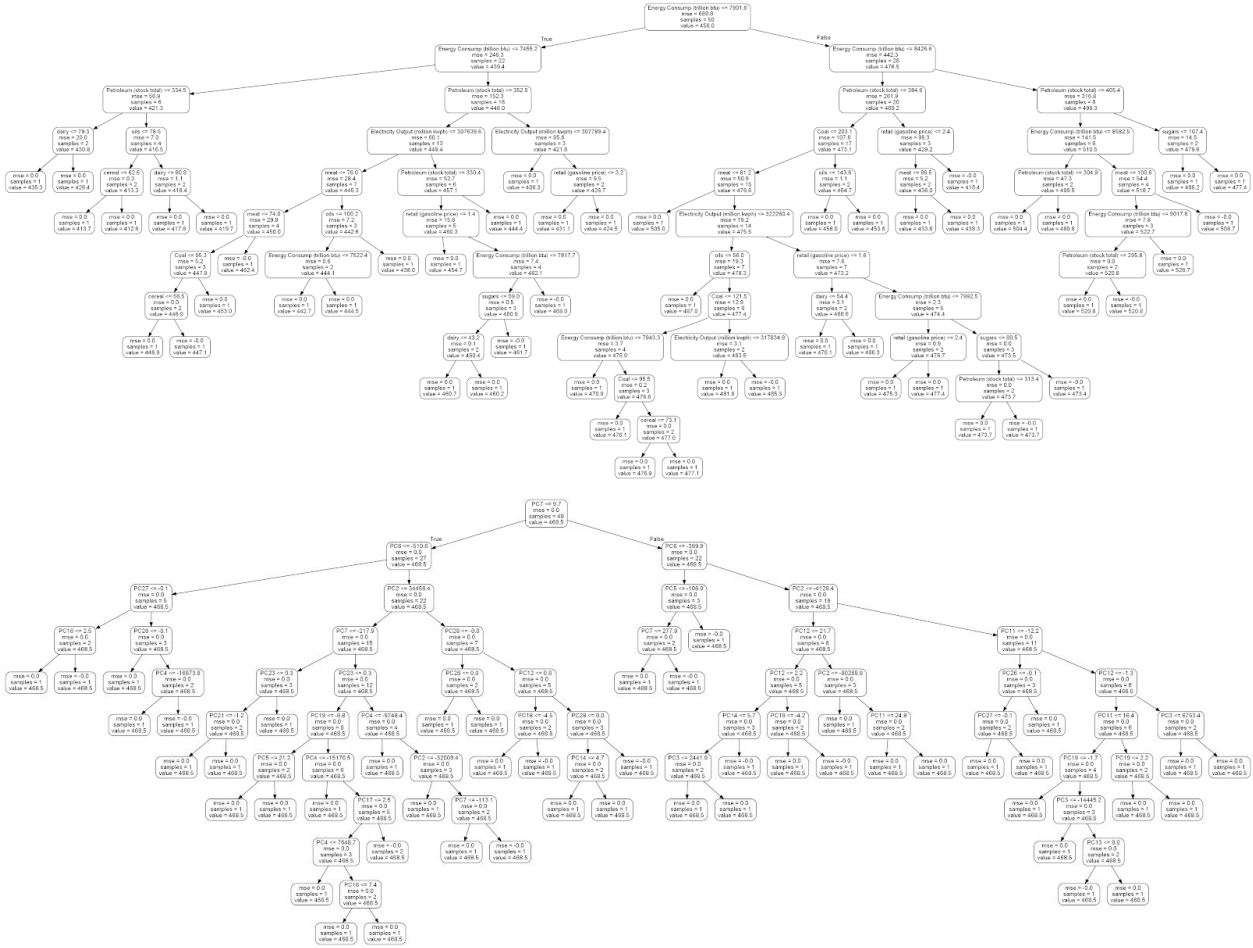
labels=np.array(y)
predictors=x1
predictor_list=list(x1.columns)
predictors=np.array(x1)
rf=RandomForestRegressor(n_estimators=1000,random_state=1)
rf.fit(x_train,y_train)
MSE=mean_squared_error(y_test, rf.predict(x_test))
print("MSE: {}".format(MSE))
importance=list(rf.feature_importances_)
importance_variables=[(feature,round(importance,6)) for feature,importance in zip(predictor_list,importance)]
importance_var_sorted=sorted(importance_variables,key=lambda importance: importance[1],reverse=True)
for pair in importance_var_sorted:
    print(str(pair[0]) + " with importance: " + str(pair[1]))

from sklearn.tree import export_graphviz
import pydot# Pull out one tree from the forest
tree = rf.estimators_[5]# Import tools needed for visualization
from sklearn.tree import export_graphviz
import pydot# Pull out one tree from the forest
tree = rf.estimators_[5]# Export the image to a dot file
export_graphviz(tree, out_file = 'tree.dot', feature_names = predictor_list, rounded = True, precision = 1)# Use dot file to
(graph, ) = pydot.graph_from_dot_file('tree.dot')# Write graph to a png file
graph.write_png('tree_returns.png')

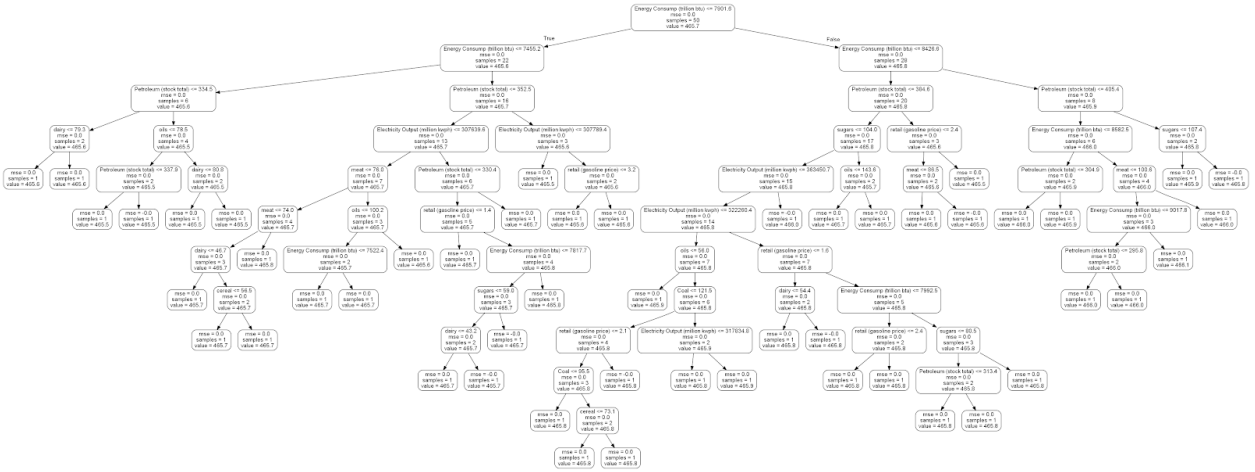
rf_small = RandomForestRegressor(n_estimators=1000, max_depth = 4)
rf_small.fit(x_train, y_train)# Extract the small tree
tree_small = rf_small.estimators_[5]# Save the tree as a png image
export_graphviz(tree_small, out_file = 'small_tree_.dot', feature_names = predictor_list, rounded = True, precision = 1)
(graph, ) = pydot.graph_from_dot_file('small_tree_.dot')
graph.write_png('small_tree_.png')
```

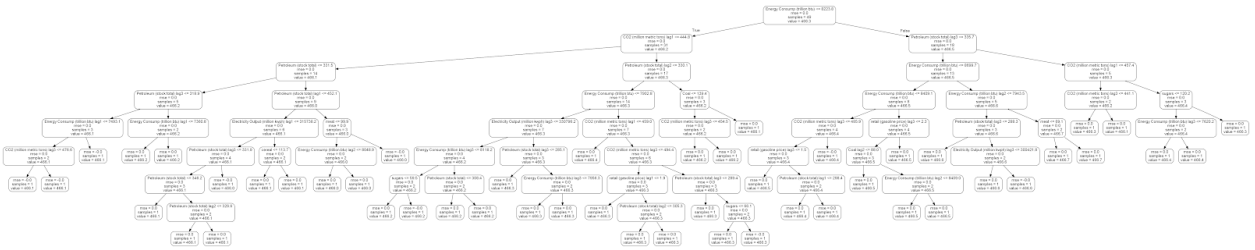
Random Forest code

Appendix 4


Energy Consump (trillion btu) <= 7901.0
mse = 469.8
samples = 50
value = 469.0
True
False
Energy Consump (trillion btu) <= 7456.2
mse = 346.3
samples = 22
value = 439.4
Energy Consump (trillion btu) <= 8426.8
mse = 442.3
samples = 28
value = 476.0
Petroleum (stock total) <= 334.5
mse = 50.9
samples = 5
value = 421.3
Petroleum (stock total) <= 352.5
mse = 152.3
samples = 15
value = 446.0
Petroleum (stock total) <= 384.5
mse = 261.9
samples = 20
value = 469.2
Petroleum (stock total) <= 405.4
mse = 316.6
samples = 8
value = 499.5
dairy <= 79.3
mse = 20.5
samples = 4
value = 430.9
oils <= 78.5
mse = 4.4
samples = 1
value = 416.5
Electricity Output (million kwph) <= 307639.0
mse = 60.1
samples = 11
value = 449.4
Electricity Output (million kwph) <= 307789.4
mse = 85.5
samples = 4
value = 421.6
Coal <= 283.1
mse = 47.5
samples = 17
value = 473.1
retail (gasoline price) <= 2.4
mse = 64.3
samples = 3
value = 429.2
Energy Consump (trillion btu) <= 8982.5
mse = 4.5
samples = 5
value = 518.5
sugars <= 107.4
mse = 1.5
samples = 3
value = 479.6
mse = 0.0
samples = 1
value = 426.4
mse = 0.0
samples = 1
value = 435.3
cereal <= 82.8
mse = 0.2
samples = 2
value = 413.3
dairy <= 80.0
mse = 1.1
samples = 2
value = 418.4
meat <= 74.8
mse = 29.9
samples = 4
value = 456.6
oils <= 100.2
mse = 7.2
samples = 7
value = 442.6
meat <= 61.2
mse = 56.9
samples = 15
value = 476.6
oils <= 183.6
mse = 1.1
samples = 2
value = 464.7
meat <= 90.6
mse = 9.2
samples = 2
value = 436.9
mse = -0.0
samples = 1
value = 418.4
Petroleum (stock total) <= 304.0
mse = 2.2
samples = 2
value = 499.5
Coal <= 102.6
mse = 54.4
samples = 4
value = 516.7
mse = 0.0
samples = 1
value = 477.4
mse = 0.0
samples = 1
value = 413.7
mse = 0.0
samples = 1
value = 412.6
mse = 0.0
samples = 1
value = 417.8
mse = 0.0
samples = 1
value = 419.7
Coal <= 95.2
mse = 5.2
samples = 3
value = 447.9
mse = -0.0
samples = 1
value = 462.4
Energy Consump (trillion btu) <= 7622.4
mse = 0.6
samples = 2
value = 444.1
mse = 0.0
samples = 1
value = 436.0
Petroleum (stock total) <= 330.4
mse = 52.2
samples = 6
value = 457.1
Electricity Output (million kwph) <= 322260.4
mse = 18.2
samples = 14
value = 475.5
retail (gasoline price) <= 1.6
mse = 7.9
samples = 7
value = 473.2
Energy Consump (trillion btu) <= 7992.5
mse = 2.3
samples = 5
value = 474.4
mse = 0.0
samples = 1
value = 529.0
mse = 0.0
samples = 1
value = 454.7
Energy Consump (trillion btu) <= 7817.7
mse = 7.4
samples = 4
value = 462.1
oils <= 56.6
mse = 19.3
samples = 7
value = 478.3
dairy = 54.4
mse = 3.1
samples = 2
value = 468.8
Petroleum (stock total) <= 295.8
mse = 9.0
samples = 3
value = 520.8
mse = 0.0
samples = 1
value = 529.7
cereal <= 59.3
mse = 0.0
samples = 2
value = 446.9
mse = -0.0
samples = 1
value = 447.1
mse = 0.0
samples = 1
value = 455.0
sugars <= 59.0
mse = 0.5
samples = 3
value = 460.9
mse = -0.0
samples = 1
value = 469.0
Coal <= 121.5
mse = 12.9
samples = 6
value = 477.4
mse = 0.0
samples = 1
value = 470.1
mse = 0.0
samples = 1
value = 486.3
retail (gasoline price) <= 2.4
mse = 0.9
samples = 2
value = 479.7
mse = 0.0
samples = 1
value = 520.0
mse = -0.0
samples = 1
value = 520.8
dairy <= 43.2
mse = 0.1
samples = 2
value = 460.4
mse = -0.0
samples = 1
value = 461.7
Energy Consump (trillion btu) <= 7643.3
mse = 3.7
samples = 4
value = 475.9
Electricity Output (million kwph) <= 317834.8
mse = 3.1
samples = 3
value = 483.5
sugars <= 80.5
mse = 0.6
samples = 3
value = 473.5
mse = 0.0
samples = 1
value = 460.7
mse = 0.0
samples = 1
value = 460.0
mse = 0.0
samples = 1
value = 476.3
Coal <= 95.5
mse = 0.2
samples = 3
value = 476.6
mse = 0.0
samples = 1
value = 481.8
mse = -0.0
samples = 1
value = 485.3
mse = 0.0
samples = 1
value = 477.4
Petroleum (stock total) <= 313.4
mse = 0.0
samples = 2
value = 473.7
mse = -0.0
samples = 1
value = 473.4
mse = 0.0
samples = 1
value = 476.1
cereal <= 73.1
mse = 0.0
samples = 2
value = 477.0
mse = 0.0
samples = 1
value = 473.7
mse = -0.0
samples = 1
value = 473.1
mse = 0.0
samples = 1
value = 475.9
mse = 0.0
samples = 1
value = 477.1

PC7 <= 9.7
mse = 0.0
samples = 49
value = 468.5
True
False
PC8 <= -510.8
mse = 0.6
samples = 27
value = 466.5
PC6 <= -359.9
mse = 0.0
samples = 22
value = 469.5
PC27 <= -3.1
mse = 0.0
samples = 5
value = 469.5
PC3 <= 34458.4
mse = 0.6
samples = 22
value = 466.3
PC5 <= -108.9
mse = 0.0
samples = 3
value = 466.5
PC2 <= -4128.4
mse = 0.0
samples = 19
value = 469.5
PC16 <= 2.0
mse = 0.0
samples = 2
value = 468.5
PC28 <= -3.1
mse = 0.0
samples = 3
value = 469.5
PC7 <= -217.9
mse = 0.0
samples = 15
value = 468.5
PC28 <= -0.0
mse = 0.0
samples = 7
value = 469.5
PC7 <= 277.9
mse = 0.0
samples = 2
value = 468.5
mse = -0.0
samples = 1
value = 468.5
PC12 <= 21.7
mse = 0.0
samples = 8
value = 468.5
PC11 <= -12.2
mse = 0.0
samples = 11
value = 468.5
mse = 0.0
samples = 1
value = 468.5
mse = 0.0
samples = 1
value = 468.5
mse = 0.0
samples = 1
value = 468.5
PC4 <= -16873.9
mse = 0.0
samples = 2
value = 468.5
PC23 <= 0.3
mse = 0.0
samples = 3
value = 468.6
PC23 <= 0.3
mse = 0.0
samples = 12
value = 468.5
PC28 <= 0.0
mse = 0.0
samples = 2
value = 469.5
PC12 <= 0.0
mse = 0.0
samples = 5
value = 469.5
PC12 <= 2.2
mse = 0.0
samples = 5
value = 469.6
PC2 <= -9026.8
mse = 0.0
samples = 3
value = 469.5
PC26 <= -0.1
mse = 0.0
samples = 5
value = 469.5
mse = 0.0
samples = 1
value = 468.5
mse = 0.0
samples = 1
value = 468.5
PC21 <= -1.2
mse = 0.0
samples = 2
value = 469.5
mse = 0.0
samples = 1
value = 466.5
PC19 <= -5.3
mse = 0.0
samples = 6
value = 469.5
PC4 <= -9748.4
mse = 0.0
samples = 4
value = 469.5
mse = 0.0
samples = 1
value = 466.5
mse = 0.0
samples = 1
value = 466.5
PC16 <= -4.5
mse = 0.0
samples = 2
value = 469.5
PC28 <= 0.0
mse = 0.0
samples = 3
value = 469.5
PC19 <= -4.2
mse = 0.0
samples = 2
value = 468.5
mse = 0.0
samples = 1
value = 469.5
PC11 <= 24.8
mse = 0.0
samples = 2
value = 469.5
PC27 <= -5.1
mse = 0.3
samples = 2
value = 469.5
mse = 0.0
samples = 1
value = 469.5
PC12 <= -1.3
mse = 0.0
samples = 6
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
PC15 <= 16.4
mse = 0.0
samples = 6
value = 469.5
PC3 <= 6753.4
mse = 0.0
samples = 2
value = 469.5
PC19 <= 1.7
mse = 0.0
samples = 4
value = 469.5
PC16 <= 2.2
mse = 0.0
samples = 2
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
PC5 <= -14445.2
mse = 0.0
samples = 3
value = 469.5
mse = 0.9
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
PC13 <= 0.0
mse = 0.0
samples = 2
value = 469.5
mse = -0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5
PC4 <= 7848.7
mse = 0.0
samples = 3
value = 469.5
mse = -0.0
samples = 2
value = 469.5
PC16 <= 7.4
mse = 0.0
samples = 2
value = 469.5
mse = 0.0
samples = 1
value = 469.5
mse = 0.0
samples = 1
value = 469.5





Random Forest Diagrams