A Study of the relationship between fossil fuel investment and green technologies investment

University ID: 2286975

Abstract - This paper studies the relationship between investment in fossil fuels measured by the stock price of the top five American oil companies and the West Texas Intermediate price, and the three metals involved in the climate transition measured by the futures prices for Copper, Cobalt and Nickel. Correlations have been discovered between specific stocks and metals. Predictive regression models have also been developed that have a strong predictive potential as well.

Introduction

The most important challenge facing humanity in the twenty first century is the climate crisis. Climate change will have profound effects on humanity from micro (home valuations in coastal regions) to the macro (habitability of certain regions for example the middle east). This is why the mitigation of these effects is extremely important. To help with the mitigation a large amount of investment in battery technologies, solar panels, wind technology etc is required. More specifically these technologies require specific commodities especially certain metals. These metals are Copper, Nickel, Cobalt and Lithium.[2]

However, there is only so much investment that can occur globally. Which means there is a constraint on how much can go into green infrastructure. Even more alarming is that investment can go in the opposite direction. This can be done by investments in fossil fuel companies and oil. This is the focus of this paper. This paper asks if there is a relationship between the investment in metals required for the green transition (Copper, Nickel and Cobalt) and investment in fossil fuel resources. Namely the price of oil (West Texan Intermediate price) combined with the stock price of the top five largest (American) fossil fuel corporations. More specifically Exxon Mobil (XON), Chevron (CVX), Philips 66 (PSX), Marathon Petroleum (MPC) and Valero Energy Corporation (VLO).

Background

Fossil fuels can exist in many different forms for example Jet fuel, liquified natural gas (LNG). These fossil fuels are traded in global commodity markets for example the Chicago Mercantile Exchange (CME). This paper will focus on brent crude oil. The reasoning for this is due to crude oil being unrefined petroleum, therefore it can be used to create multiple kinds of fossil fuels through processes of refinement.

The number one fossil fuel producing country in the world is the United States.[1] The best representative of crude oil prices is West Texan Intermediate (WTI). Whenever oil prices tend to be referred to in the media, they usually are referring to WTI which is measured in dollars per barrel.

For the energy transition to green technologies a vast volume of metals will need to be purchased. This could be used for building solar panels, battery storage and other low carbon sources of energy. The 4 metals required as recognised by the international monetary fund are Cobalt, Nickel, Copper and Lithium.[2] The way those metals are currently traded are via multiple forms but the main two are Futures and spots.

Spots are metals sold on the "spot" meaning immediate delivery. While Futures are contracts that sell at a certain time and date in the future. This means that futures contracts are more of an indication of investment and long-term strategic thinking, while spots are a sign of short-term trading. This is why this paper will focus on futures contracts.

The biggest metal exchange in the world is the London Metal Exchange (LME), therefore that will be used as the primary pricing system for the futures contracts. Another benefit of using LME prices is even though it is based in London, LME is priced in dollars. This will prevent pricing distortions due to exchange rate fluctuations. (For example, the Brexit vote)[3].

However, there is a problem with measuring lithium prices. Lithium future contracts did not have representation in the LME until 2021.[6] This means that the data for lithium contracts is not as extensive as other metals. So for this paper we will remove lithium futures.

Tools

- Python - Python is a programming language mainly used by data analysts because of its simple syntax and extensive libraries. For this paper python 3.9.18 was used (inside the Anaconda framework).
- Pandas - Pandas is a python library. That is used for data manipulation and analysis. This will be used extensively for data cleaning. For this paper pandas 2.1.1 was used.
- Scikit-learn - Scikit learn is a python library used for machine learning/data mining tasks. It has many machine learning algorithms from supervised linear regression to unsupervised algorithms such as K Nearest Neighbour. It also has metrics built in to measure the performance of those algorithms. For this paper version 1.3.0 was used.
- NumPy - NumPy is a python package that allows for efficient calculations on arrays. For this paper

version 1.26.0 was used.

- SciPy - SciPy is a python package that allows statistical analysis on datasets for example calculating T-Distributions. For this paper version 1.1.3 was used.
- Seaborn - Seaborn is a data visualisation tool that can plot graphs for example a scatter plot. For this paper version 0.13.0.

The data

The datasets that will be used in this paper will be an oil index and the price of futures contracts for the metals. Before delving into what each column means there needs to be an understanding on why we chose this form of data.

For all the data there is a focus on daily pricing information. For example, closing price. The reason for this is due to this being an exploration of investment rather than trading. So, we care about long term trends rather than minute by minute price movements. This also reduces the number of outliers in the data set due to malfunctions in the markets. An example case being flash crashes.[8]

Firstly, the features of the oil index are required:

Fossil Fuel Index

- Date - The date of the record of data.
- PhilipsClosing - PhilipsClosing is the price of the Philips66 closing price (price of stock at the end of the day). This is measured in United States Dollars.
- ExxonClosing - ExxonClosing is the price of the Exxon closing price (price of stock at the end of the day). This is measured in United States Dollars.
- ValeroClosing - ValeroClosing is the price of the Valero closing price (price of stock at the end of the day). This is measured in United States Dollars.
- MarathonClosing - MarathonClosing is the price of the Marathon closing price (price of stock at the end of the day). This is measure in United States Dollars.
- Price - Price of WTI oil at that date. This is measured in United States Dollars.

The source of the WTI crude oil prices is the St. Louis Federal Reserve data repository (FRED). [7] FRED is a source of economic data maintained by the St. Louis Federal reserve and is widely used by economists. The original source of the WTI price is the U.S. Energy information administration.

In addition, the stock data was retrieved from Yahoo finance.[4] Yahoo finance is data provider is ICE Data Services. [9]

Green Metals

Cobalt:

- Date - The date of the record of the data.
- CobaltPrice - The closing price of cobalt futures on that specific date.

Nickel:

- Date - The date of the record of the data.
- NickelPrice - The closing price of nickel futures on that specific date.

Copper:

- Date - The date of the record of the data.
- CopperPrice - The closing price of copper futures on that specific date.

The source of the metals data is investing.com. [5]

Data cleaning

Since the data comes raw from multiple sources a data cleaning process is required. (Note: the data was accessed on the 2023-12-29).

Oil Index

Stock Price Data:

Firstly, a standardisation of the data is required. When receiving the unclean data, the following columns in the CSV file were available: Date, Open, High, Low, Close, Adjusted close and Volume. The first piece of the data clean-up was removing columns that are not required. The columns that were removed were adjusted close and volume. Volume due to it not really being related to the actual value of the stock. This is due to the volume of stocks not being related to actual investment but rather if the company has decided to buy back their own shares or issue new stocks for business reasons. 'Adjusted close' was removed because again this is more to do with corporate action rather than actual investment.

Next, was a standardisation of the dates of all these stocks to match other pieces of data. This was done with Panda's python package which can convert the string data to a date time object. After that then the conversion of that date time object to just a date object is required.

The next step is to constrict the number of records to dates that have records in all the different stocks. Exxon, Chevron, Marathon, Philips66 and Valero have

record start dates of 1962-01-02, 1962-01-02, 2011-06-24, 2012-04-12 and 1982-01-04 respectively. This implies that the start date would most likely be 2012-04-12 because it is the latest start date. To determine which is in all the other stock record another panda's selection tool will be used. The function isin will be used to determine which data is in the other 4 stocks using date as the main index. This leaves 2936 records of data in the dataset for each stock.

Finally, the data is merged on the date so there are the columns are: Date, OilPrice, PhilipsClose, ValeroClose, ExxonClose, ChevronClose, Marathon Closing Price.

West Texas intermediate:

The West Texas Intermediate oil price only provides two columns, namely Date and DCOILWTICO (oil price in Dollars).

The first task that needed to be done is a standardisation of the date. This will use the same techniques as the stock price, so I will not repeat the flow. The same for selecting what data is matching both the stock prices and WTI. Another aspect that needs to be worried about in some cases at certain dates there is a "." Instead of a price. This is due to American holidays occurring on those days. For example, 2023-02-20 is Presidents Day.

Next, we use the dates to see which match with the cleaned stock data and remove the rest of the data. Finally, we merge the column to the stock data to create an oil index. This adds a column of oil price. Finally, a removal of outliers is required. The main outlier case in the WTI dataset happened during the early days of the covid-19 pandemic where the price went negative price.

Green Data

Metals

The futures data comes from the same source therefore the data has the same columns. Therefore, the cleaning of all three metals will be the same. The columns provided are Date, Price, Open, High, Low, Vol. and Change

Firstly, we drop the Open, High, Low, Vol. and change % because they are not relevant to the prediction. Next, we standardise the date using pandas just like the oil index. Finally, the removal of data not in the oil data sets. This will be unique for each metal because they come from different markets (therefore there are three datasets Y). A quirk of this data source which is there is a comma indicating a thousand but a quick removal using the panda's python package solves this problem.

This means there are three sets of X's ($\{X_{\text{Oil Cobalt data}}, X_{\text{Oil Copper data}}, X_{\text{Oil Nickel data}}\}$) and three sets of Y's ($\{Y_{\text{Cobalt futures prices}}, Y_{\text{Copper future Prices}}, Y_{\text{Nickel future prices}}\}$) in the learning problem.

Hypothesis

The hypothesis this paper proposes is that there is a correlation between the investment in fossil fuel resources and in green metal futures. This paper proposes this hypothesis because when fossil fuel prices go up in value then the cost of energy increases. This cost of energy increase incentivises industries and speculators to find alternative cheaper forms of energy. These alternative cheaper forms of green energy will require the infrastructure investment in certain metals (copper, cobalt and nickel). Thus, the price of these metals will also increase at the same time.

Data analysis

Correlations

When analysing the data sets and preparing them for regression, the first test is to check if the X values have any correlations with the Y values. This is far more difficult for a multi-dimensional X as it is hard to both visualise and analyse. So, in this paper we will simply check if there is a correlation feature by feature. The technique we will use to observe these correlations is the Pearson coefficient (r). This is defined as

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Where $\bar{X}$ and $\bar{Y}$ means the value of the average of both X datasets and Y data sets:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Where n is the number of samples.

We are using the population correlation coefficient this is due to there being enough samples such that there is not much of a deviation between the population correlation coefficient and the sample correlation coefficients.

Correlation coefficients have a range of [-1,1], where if ρ tends to -1 the higher the negative correlation and if rho tends to 1 the higher the positive correlation (correlation meaning if one variable increases the other variable increases to), while if rho is closer to 0 there is no correlation.

In addition, when studying the correlations, the search for correlations will include a re-framing of the data, in Log(Metal price data) vs Log(Fossil fuel data), Log(Metal price data) vs Fossil fuel data, and Metal price data vs Log(Fossil fuel data). (When referring to Log the base is 10).

Cobalt

The correlations discovered in the Cobalt price to stock would indicate that there was no strong correlation to any stock outside the Valero closing price.

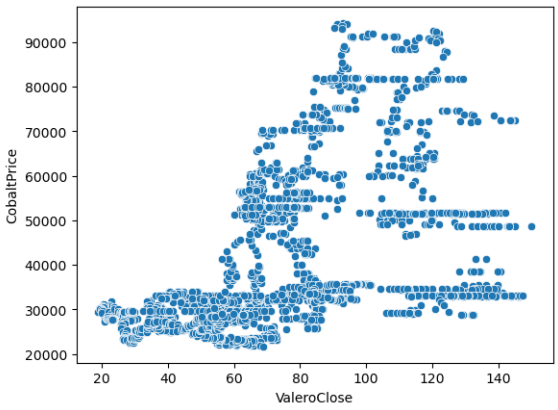

Figure 1: Cobalt Futures price vs Valero Closing Price

Which had a correlation of 0.387. Exxon had the worst correlation coefficient output of -0.079. As seen through figure 2.

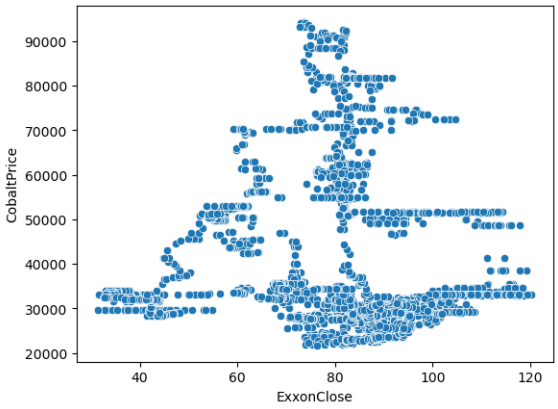

Figure 2: Cobalt Futures price vs Exxon Closing Price

To try to re frame the data and improve the correlation coefficients I took the log/log and the log/price to see if that would significantly change anything. For most of the stocks the changes were not significant except for the case of Marathon where and improvement from 0.347 to 0.494(Log/Log) and 0.457(log vs cobalt). This only occurred if the square of the Valero stock was had a logarithmic transformation.

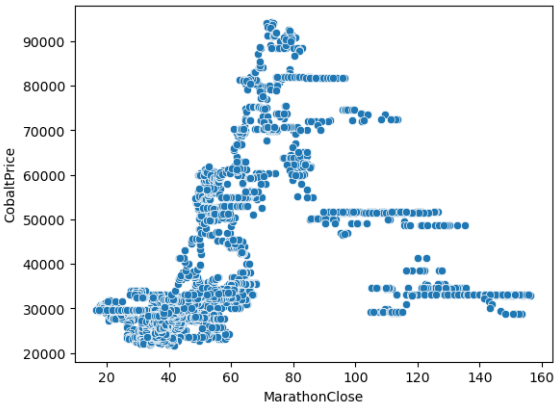

Figure 3: Log(Marathon closing price) vs Cobalt Futures price

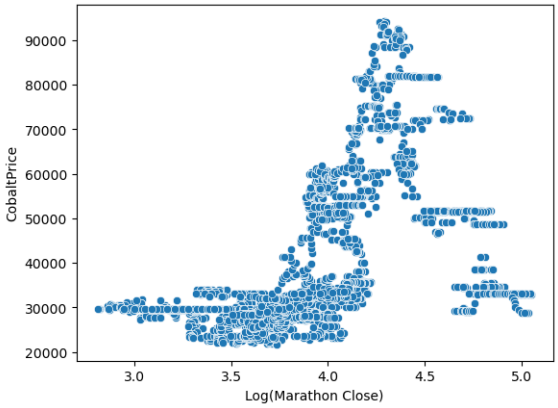

Figure 4: Log(Marathon closing price) vs Cobalt Futures price

The case is also far worse in the relation between WTI and Cobalt prices, no correlations could be found that are larger than 0.2 (even with the log/log graphs).

The insights that someone can take from this are when building a model, they can take the log of the Valero stock instead of the log of just taking the Valero stock because it has significantly higher correlations. This will improve the predictive capacity of said model.

Nickel

The correlations discovered in nickel prices were far more promising. With PhilipsClose there was a low level of correlation in all frames (highest being 0.122). A similar situation occurred with ExxonClose with a low correlation (Highest being 0.269).

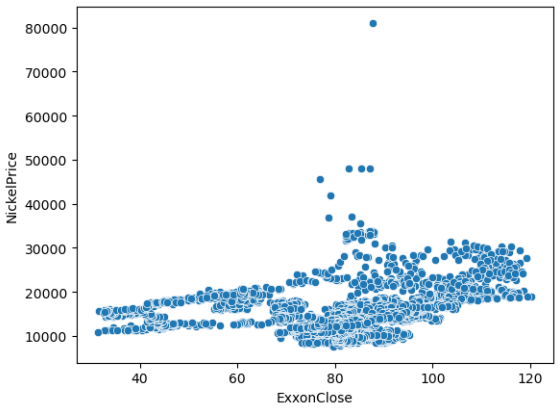

Figure 5: Nickel Futures Price vs Exxon Closing Price

However, the rest of the stock prices alternated between moderate and high correlation. ValeroClose has a correlation of 0.417 (at the maximum) and Marathon has a maximum correlation of 0.521 throughout all the frames.

The most promising correlation was ChevronClose and Nickel prices. This is due to the correlation coefficient being 0.7055. This indicates that the price of chevron

stock and nickel prices are highly correlated with each other.

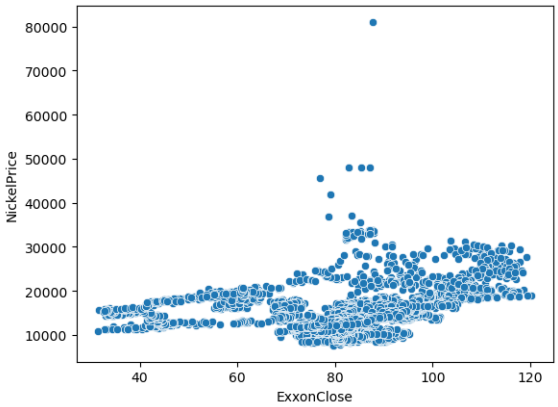

Figure 6: Nickel Futures Price vs Exxon closing price

Another promising aspect of nickel prices are their correlations with WTI. With a high of 0.685 in a log log graph.

The conclusions that will be taken from this is that no changes to the data will be required as there is not a significant improvement.

Copper

Copper was more like cobalt, weak to medium correlations. With the highest correlated feature amongst the stock data being ChevronClose (0.534 (LogLog)) but there wasn't that much difference between the changes of frames. Although the correlation of copper prices to WTI was far stronger at (0.661).

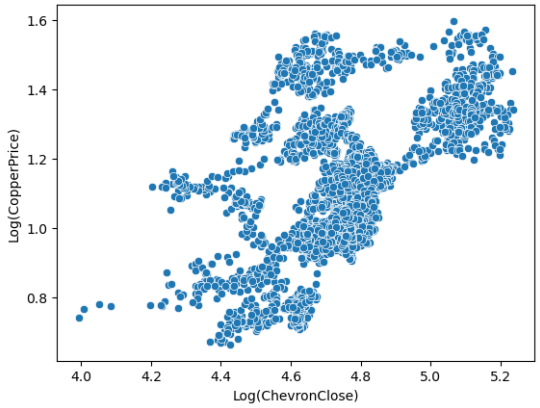

Figure 7: Log(Copper Futures Price) vs Log(Chevron Close)

The conclusions taken from this would just be feed the logarithmic version of WTI into the linear models.

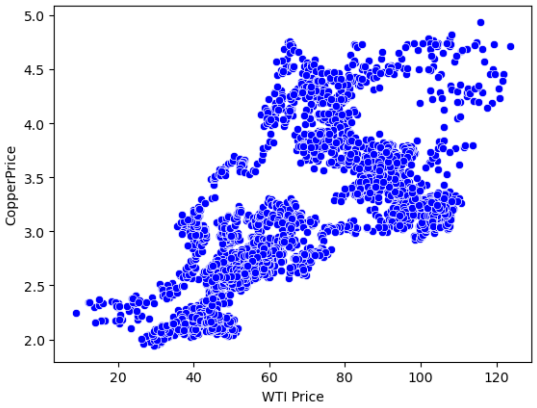

Figure 8: Copper Price vs West Texan Interchange Price

Statistical Significance

One check that must be implemented for these correlation coefficients is to verify if this is just a random event or are these statistically significant enough to be considered legitimate. The way to check this is via a hypothesis test. To perform a hypothesis test we need multiple components:

- Null hypothesis. (H_0)
- Alternative hypothesis. (H_1)
- Significance levels. (α)

In our case we are testing if the correlations are occurring by happenstance or are these correlations statistically significant enough to make some inferences about them. Therefore, the null hypothesis would be $\rho = 0$. Meaning the base case is there is not a correlation, and the alternative hypothesis is that there is a correlation namely $\rho \neq 0$. This will be a two tailed test.

We will go through the example of the correlation discovered in Chevron close and Nickel prices:

In the case on Nickel vs ChevronPrice: This test has 2871 degrees of freedom because it has 2872 observations in the sample. Next is to calculate the lower critical values and the upper critical values. In this case being -1.96 and 1.96. If the test statistic T falls out of this range the null hypothesis is rejected and if it falls into this range the null hypothesis is maintained. T is calculated by:

$$t = \frac{\rho}{\sqrt{\frac{1-\rho^2}{n-2}}} = \frac{0.7055}{\sqrt{\frac{1-0.7055}{2872-2}}} \approx 53$$

Since 53 is out of our range the null hypothesis is rejected. This was observed throughout all the correlations.

To conclude the data analytics correlations were discovered throughout every comparison. The insights that can be taken into the model formation is that cobalt had the weakest correlations overall, followed by copper than nickel. Exxon had the weakest correlation of all the stocks and might be adding noise to any

models and finally all these correlations have statistical significance.

Selecting and applying models

Models

To inspect if there are any relationships between fossil fuel investment and green investment multiple models were developed. These models were regression models namely, linear regression, Lasso regression and Ridge regression.

Firstly, Linear regression is a model that finds the "line of best fit" for a relationship of X to Y. Due to the data we are using being multi-dimensional the input variable is a matrix, and the coefficients that are applied are also a matrix.

Next, we use ridge regression which takes a hyper-parameter α which is defined by the user. Linear regression has a tendency of over fitting to the data, therefore there is a penalty added to the model to prevent this from happening. The penalty is α.

Finally, Lasso regression tries to solve the issue in a different way. It also takes in a hyper-parameter of α as defined by the user. However, the model is constrained from over fitting in a different way.

Quality of regression

The metric which is being used to measure the quality of the regression would be R-squared value. R-Squared is a quotient of the residual sum of squares and the total sum of squares defined as follows:

$$R^2 = 1 - \frac{SS_{\text{reg}}}{SS_{\text{tot}}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where $\hat{y}$ is a prediction and $\bar{y}$ is the mean. The range for this function is [0-1]. Therefore, we will define 0.6 and above as low variability.

Searching for best parameters

Searching Techniques: To prevent over fitting a 5-fold cross-validation was used. This meant that 4 folds were for training and 1-fold was for testing. This made sure that the model was constantly adapting to new data on each training iteration. Then the data was tested on the remaining fold. After that the R-Squared value was calculated and finally the mean of those 5 R-squared values were taken.

Finally, the best model (highest R-Squared) was taken and checked against the whole data set. (Those are the values this paper will use).

Tuning Hyper-parameters: For both α's for ridge and lasso regression were exhaustively searched through. This was from 10^{-10} to 10^{10} going up by 1 in power.

Application of models

For the linear regression model, moderate levels of variability $R^2 = 0.452$. This was expected as in the previous section there wasn't any large correlations discovered.

Cobalt

For the linear regression model, moderate levels of variability with a $R^2 = 0.452$. This was expected as in the previous section there wasn't any large correlations that were detected.

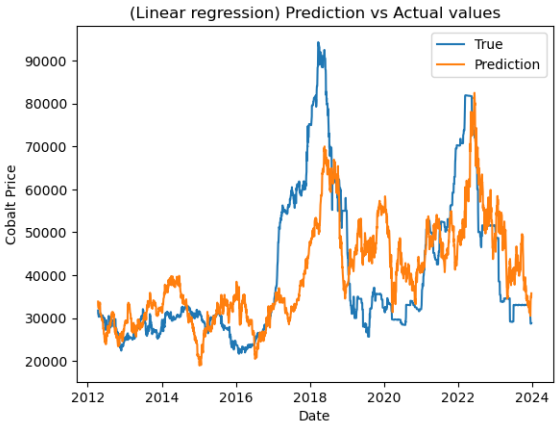

Figure 9: Cobalt futures prices vs Copper futures prediction

This is also true for ridge search with the best hyper-parameter $\alpha = 10^2$ and a $R^2 = 0.452$. Which indicates there is not a high level of overfitting in the Cobalt linear regression and would imply that Lasso would not make much of a difference. Which is observed in the Lasso performance, where the best hyper parameter was 10^1 and a R^2 value of 0.453.

Nickel

The results from the nickel data were far more promising. The R-Squared value had vastly better results for linear regression with a value of 0.7519. This means that there was a strong predictive capacity for nickel prices from the oil index. This can be deduced from nickels correlation with Chevron. Also there must be hidden correlations that were not picked up in higher dimensional planes that were not detected before.

This benefit has been seen in the Lasso regression, but the performance improvement has been negligible. The hyper parameter in this case is $\alpha = 10^1$ and a R-Squared of 0.7519. This also implies over fitting is not occurring as when the L2 penalty is being applied. They seem to be converging on the same coefficients.

Finally, Ridge regression is showing similar behaviour as the other two algorithms. It's $\alpha = 10^3$ while its R-squared is 0.7518. The conclusions one can gather from this is Copper prices are highly correlated with

the price of oil and this model can have some predictive value.

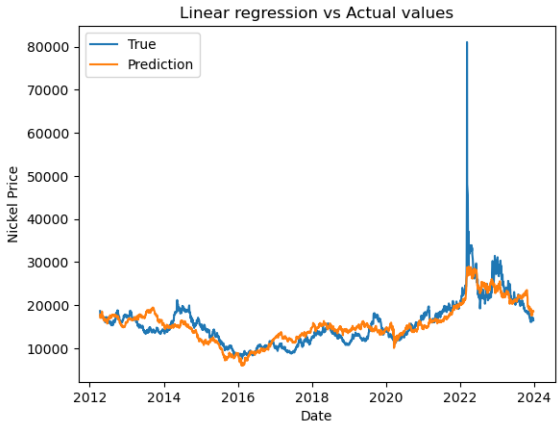

Figure 10: Nickel Futures price vs Nickel Futures prediction

Copper

The results from the copper data were the most surprising. This is due to the high performance of each regression. Which provides the implication that there is some hidden correlation in a higher dimension that was not picked up. With an R-Squared (0.789) value nearly surpassing the most promising case for nickel for regression being 0.7519. It is a similar story for both Lasso (with α being 10^1 and the same R-Squared to 3.S.F) and Ridge (with α being 10^{-10} with the same R-Squared to 3.S.F) regression. There is no overall improvement in the different models with penalties, therefore over fitting was not an issue.

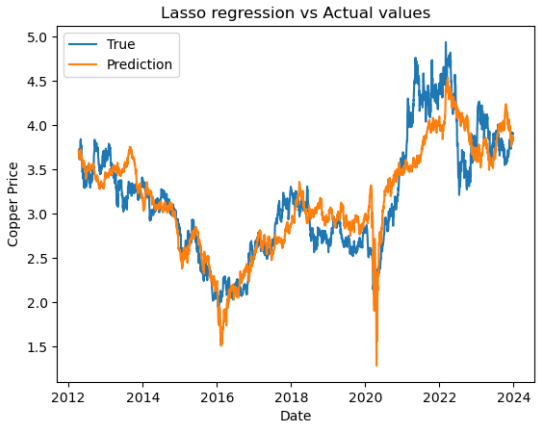

Figure 11: Copper Futures Price vs Copper Futures Prediction

Conclusions and Future works

To summarise, in this paper the relationship between climate investment and fossil fuel investment has been explored. This has yielded results stating there are linear correlations. These relationships have also been proven to be statistically significant through T-Distribution tests. With the strongest relationship being cobalt and nickel. Another outcome of this paper is the predictability of metal futures when taking all these fossil fuel investment indicators in aggregate. This has implied a correlation within the fossil fuel indicators in a higher dimension. This information can be used to help inform policy decisions for the climate transition, namely by increasing oil prices which could lead to an increase in green energy alternatives.

As for future work, more data could be added to the oil index. This could include global oil companies and oil markets for example BP and the Dubai mercantile exchange. This will allow for a more global view of the relation between fossil fuel investment and the transition metals. Another piece of work that can be explored is the long-term predictive value. Since this paper has proven that there are correlations between transition metal investment and oil investment. A more complex model can be used for future forecasting, for example Time Series analysis using a LSTM model. Final project that can be undertook is the expansion of the domain dataset by adding natural gas. Natural gas is being thought of as a bridge fuel and is becoming more predominant in the fossil fuel sector. Therefore, the relation between natural gas investment and green investment must be studied.

References

[1] U.S. Energy Information Adminstration. Weekly u.s. field production of crude oil (thousand barrels per day). `https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=WCRFPUS2&f=W`, 2024. [Online; accessed 09-January-2024].

[2] Lukas Boer, Mr Andrea Pescatori, and Martin Stuermer. *Energy transition metals*. International Monetary Fund, 2021.

[3] Thong M. Dao, Frank McGroarty, and Andrew Urquhart. *The Brexit vote and currency markets*, volume 59. 2019.

[4] Yahoo Finance. Yahoo finance – stock market live, quotes, business finance news. https://uk.finance.yahoo.com, 2024. [Online; accessed 09-January-2024].

[5] Investing.com. Investing.com - stock market quotes financial news. https://www.investing.com, 2024. [Online; accessed 09-January-2024].

[6] London metal exhange. About lithium — london metal exchange. https://www.lme.com/Metals/EV/About-Lithium, 2024. [Online; accessed 09-January-2024].

[7] St. Louis Federal Reserve. Crude oil prices: West texas intermediate (wti) - cushing, oklahoma (dcoilwtico) — fred — st. louis fed. https://fred.stlouisfed.org/series/DCOILWTICO, 2024. [Online; accessed 09-January-2024].

[8] U.S. Security and Exchange Commission. Findings regarding the market events of may 6, 2010. https://www.sec.gov/news/studies/2010/marketevents-report.pdf, 2024. [Online; accessed 09-January-2024].

[9] ICE Data Services. Ice. https://www.ice.com/index, 2024. [Online; accessed 09-January-2024].