R_project

Phillip An, Paul Park, Min Jin Kang

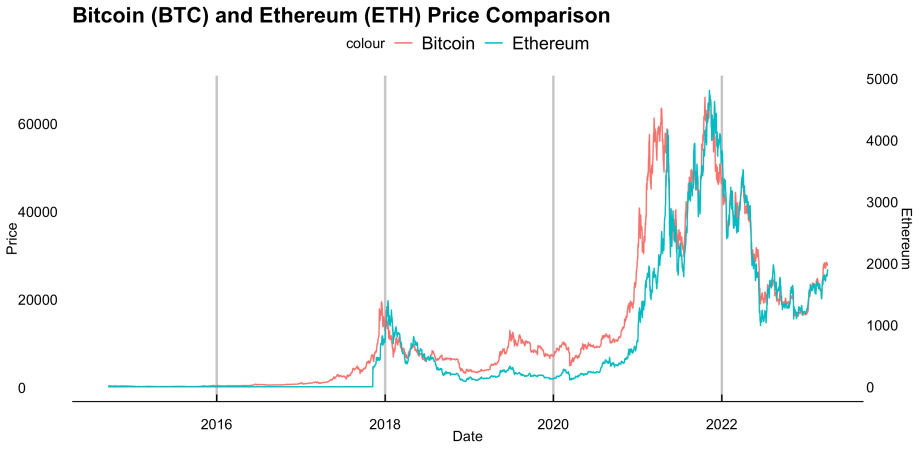
2023-04-05

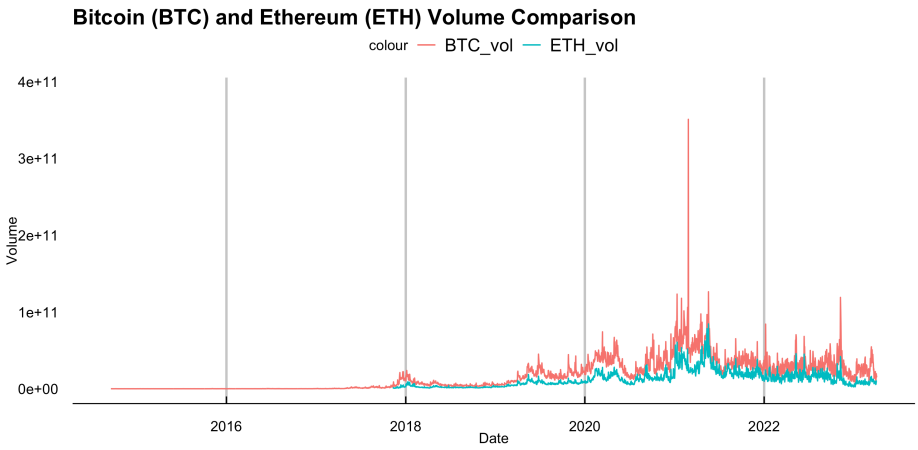
Abstract

This paper is a study of which factors have a large impact on cryptocurrencies. Bitcoin and Ethereum were chosen as representative cryptocurrencies for our research, as these are two of the most popular and widely used with a significant market capitalization and user base. This study utilizes a random forest model to analyze the relationship between predictor variables and an outcome variable. Variable importance plots are used to identify the most significant predictor variables, and (meaningful conclusions) are drawn from the analysis.

Introduction

Bitcoin, a type of digital currency run on blockchain technology, has recently gained traction as a potential currency to substitute for fiat money such as the U.S. dollar. A massive 'money-printing' by the Federal Reserve in response to the 2008 financial crisis and the Covid crisis has raised concerns about the strength and sustainability of the dollar value. Since then, market participants have been increasingly treating Bitcoin as a hedging asset. Devoid of the intrinsic value, however, Bitcoin has experienced a massive fluctuation in terms of asset prices. Specifically, at the onset of the Covid crisis in March 2020, the Bitcoin price dipped as low as `$5,165`. It grew more than tenfold to `$61,283` per Bitcoin in exactly one year, and hit as high as `$64,400` before crashing back down to hovering around `$25,000` in the beginning of the year 2023. Therefore, investors have taken interest in predicting the short-term Bitcoin price. → (too many have pp's)





This report answers 3 questions. First, we attempt to test a variety of models and find a model that best predicts the Bitcoin price. We test the prediction power between CART, Random Forest, and Gradient Boosted trees model. We will pick the model that gives the lowest value of RMSE. Second, after identifying the best-predicting model, we attempt to find a variable that contributes the most to the price prediction. Next, we will run the model not only on Bitcoin but also on Ethereum, the next popular cryptocurrency, using common features. In doing so, we are able to identify which features matter more to each cryptocurrency. This finding can help investors make well-informed investment decisions; they may be able to diversify their digital asset portfolios in response to technical or macroeconomic shocks using our model. Specifically, S&P 500, the market performance index which include the top 500 companies in terms of market capitalization sizes, contribute to the price changes of both cryptos the most. However, we find that gold price change affects Bitcoin price change more, and oil price change affects Ethereum price change more.

Methods

Data:

After reading a few existing studies regarding price prediction of an investment asset, we have learned that a wide spectrum of macroeconomic and market performance factors, such as inflation, interest rates, and market volatility are incorporated in building price predicting models. In addition to these common factors, we have also added some other variables we deemed important in terms of predictive power. Our data set consists of relevant daily asset prices, macroeconomic and market performance indicators, which are mainly collected from Yahoo Finance and Federal Reserve Economic Data (FRED).

First, we have daily prices of Bitcoin (BTC), Ethereum (ETH), oil, and gold. Inflation is measured by two proxies, yield on 10-year Treasury note (TNX) and 13 Week Treasury Bill (IRX). CBOE volatility index (VIX) estimates equity market volatility, while CBOE crude oil volatility index (OVX) measures oil market uncertainty. SP500, which tracks the top 500 U.S. stocks, is used to compute the US stock market performance. All of these data were collected from Yahoo Finance.

In addition to aforementioned inflation proxies, 5-Year breakeven inflation rate (inf5y), which implies market participants' inflation expectation for the next five years, was also added. Equity Market Volatility: Infectious Disease Tracker (DISEASE) was included to account for the economic impact of COVID-19. U.S. dollar index (DXY) measures the performance of dollar against a basket of other world currencies. Policy-related uncertainty is measured by Economic Policy Uncertainty Index (EPU), and stock market uncertainty is measured by Equity Market Uncertainty Index (EMU). All of these data were collected from FRED.

Lastly, we added Credit Default Swap (CDS), which basically is a financial derivative through which a seller can swap his credit risk with that of a buyer. As this indicator measures dwindling of centralized financial

markets, we thought it would be interesting to see its relationship with the price action of cryptocurrencies, a decentralized asset class. This data was separately collected from Investing.com.

sumtable {vtable} Summary Statistics

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
DXY	2304	112	7.3	93	110	116	128
CDS	1969	11	12	2	5.5	12	86
EMU	3381	70	94	4.8	13	91	1230
EPU	3381	123	98	3.3	64	144	861
Inf5y	2315	1.8	0.54	0.14	1.5	2.1	3.6
Disease	3380	5.8	10	0	0	8.4	113
Oil	2332	63	21	-38	48	74	124
Gold	2330	1454	273	1051	1240	1753	2052
Bitcoin	3123	13175	16045	178	715	19048	67567
Ethereum	1974	1153	1164	84	217	1744	4812
IRX	2329	0.95	1.2	-0.1	0.043	1.7	4.9
TNX	2329	2.2	0.75	0.5	1.7	2.7	4.2
VIX	2331	18	7.5	9.1	13	22	83
OVX	2331	40	20	14	29	46	325
SP500	2331	2897	844	1742	2105	3644	4797
Month	3384	6.4	3.5	1	3	9	12

Methodology:

In predicting Bitcoin and Ethereum price, we compare three main models: Classification and Regression Trees (CART), Random Forest and Gradient Boosting. To briefly recap what we learned in class, tree is a simple predictive model that is widely used in machine learning. **CART**, also called "recursive partitioning", is a basic tree-fitting algorithm. Basically, we grow the tree recursively as to make deviance as small as possible. When we reach our minimum size or complexity stopping points, we will stop growing and prune back to make candidate trees. Lastly, we will choose via cross validation (min or 1SE).

Random forest is perhaps the most popular generic nonparametric regression technique as the model not only requires little to no cross validation and is also fast and effective. Here, we will fit trees to number of bootstrapped samples of the original data. This process, also called bagging, usually produces a better fit with lower variance than a single tree. It adds more randomness as we 'randomly' choose features subsets in building a tree, hence the name 'random forest.' After fitting a tree to each bootstrapped sample, we will average the predictions of all the different trees, producing an aggregated result, which should be more accurate.

Gradient boosting is an ensemble method like random forests. However, here you recursively fit simple trees to its 'residuals'. That is, while random forests fits trees simultaneously, gradient boosting builds one tree at a time. This model adds the newly crushed tree into the fit in each stage along the way and so the final fit is the sum of many trees. Gradient boosting can work better than random forests with finely-tuned parameters. However, it is more sensitive to noise, thus more easily encounters over-fitting problems.

Our random forests give us **variable importance plots**, which provide a list of the most significant variables in descending order. Using these plots, we can gain some useful insight into which variables contribute the most to our model. We compare the results for Bitcoin and Ethereum.

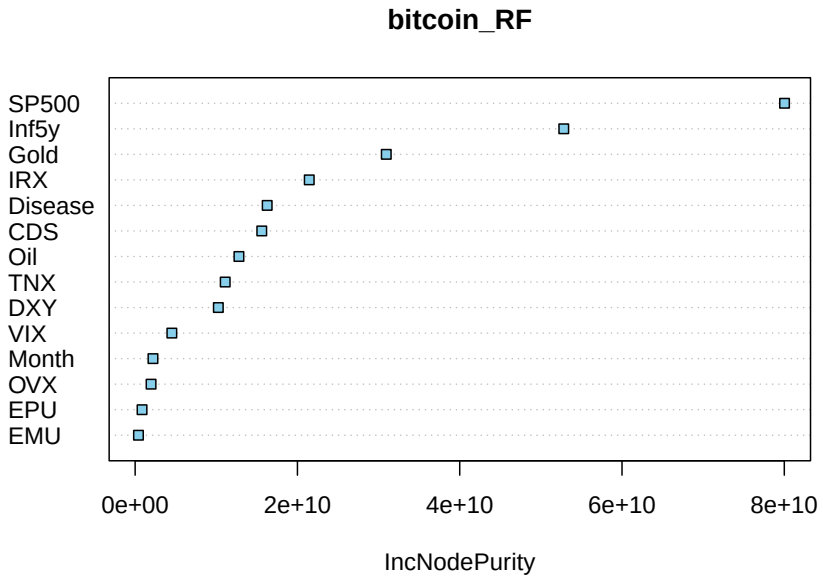
After our model has been fit, we calculate **partial dependence plots**. These plots visualize the relationship between price and the five most significant variables from the variable importance plots, taking account of the joint effect of other features.

Results

Bitcoin

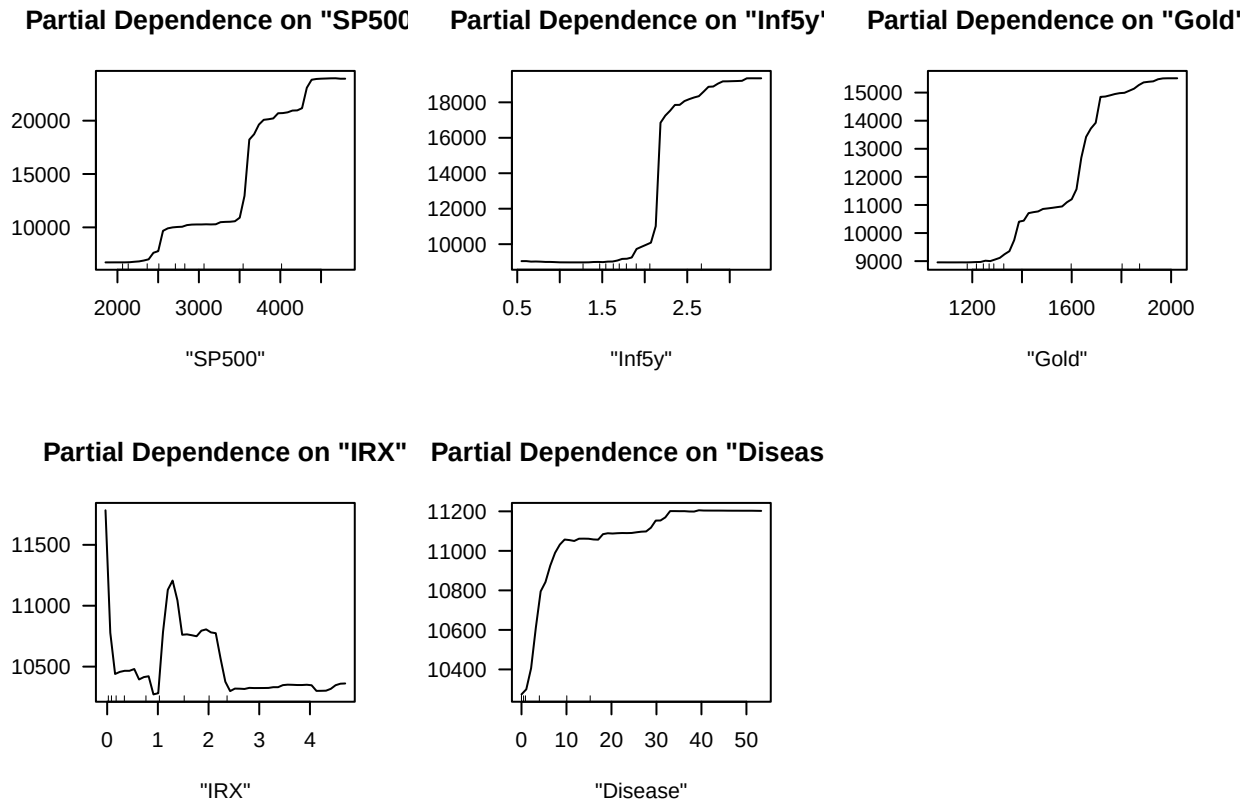
We used CART, Random Forest, and Gradient Boosted trees model and compared out-of-sample RMSEs, and we could check that the Random Forest is the best performance on the testing data. `Bitcoin` is the target variable, and the rest of the variables, excluding the `DATE` variable, are used as predictors. We used the randomForest function to fit a model and used the VarImpPlot function to display the variables which highly contribute to the model.

Random Forest



We could check `SP500`,`Inf5y`,`Gold`,`IRX`,`Disease` are top 5 important variables for bitcoin.

Below is the partial dependence plots to isolate the partial effect of specific features on the outcome. Partial dependence plot is a method used to analyze the relationship between the target variable (dependent variable) and a specific predictor variable while holding all other predictors constant.



All variables, except for `IRX`, shows an increasing dependence plot. We can interpret `SP500`, `Inf5y`,`Gold`,`Disease` features have a positive effect on predicted outcome, and `IRX` has a negative effect on predicted outcome.

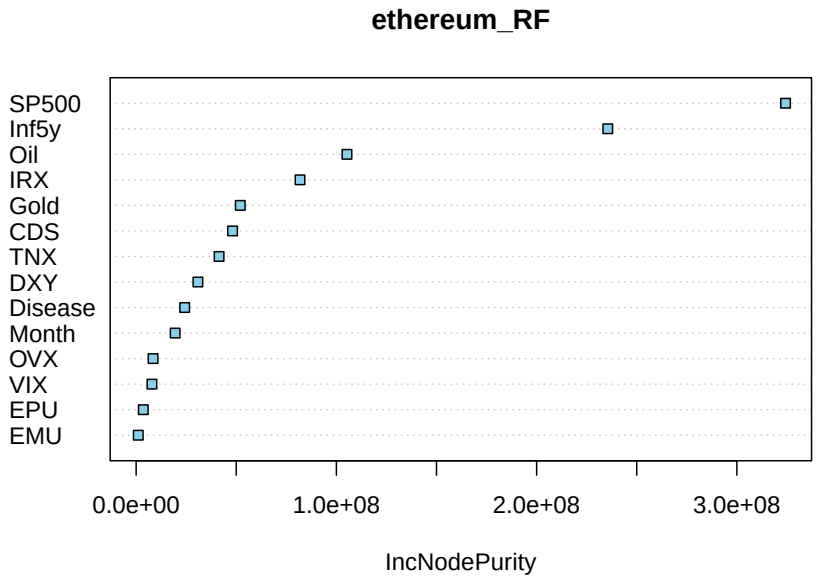
Table 1: Model performance with out-of-sample RMSEs (Bitcoin)

Model	RMSE
CART	2295.453
Random Forest	1076.176
Gradient Boosting	1283.753

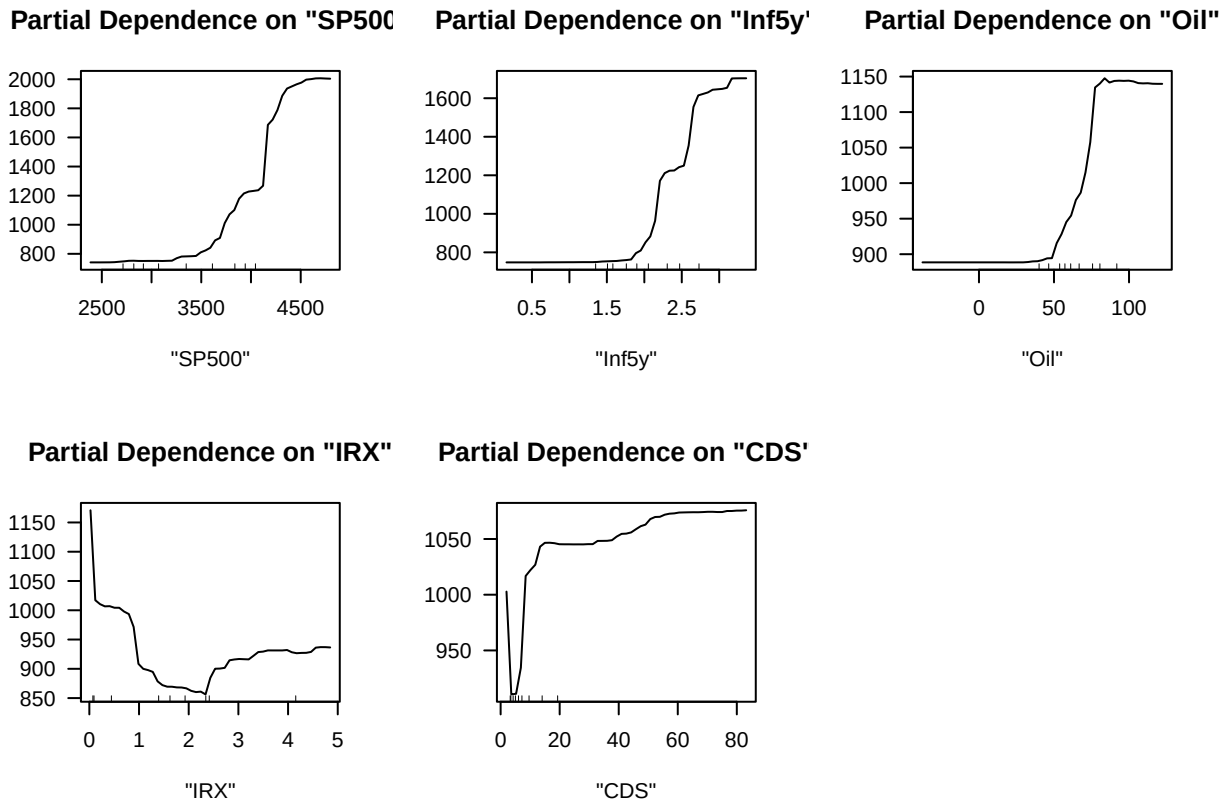
Ethereum

Same as Bitcoin, to predict Ethereum, we used CART, Random Forest, and Gradient Boosted trees model and compared out-of-sample RMSEs, and we could check that the Random Forest is the best performance on the testing data. `Ethereum` is the target variable, and the rest of the variables, excluding the `DATE` variable, are used as predictors. We used the randomForest function to fit a model and used the VarImpPlot function to display the variables which highly contribute to the model.

Random Forest



We could check `SP500`,`Inf5y`,`Oil`,`IRX`,`CDS` are top 5 important variables for ethereum. `SP500` and `Inf5y` seem to have the highest importance in both bitcoin and ethereum, but it is an interesting result that there is a difference that gold has a great influence on bitcoin and oil has a great influence on ethereum. Below is the partial dependence plots to isolate the partial effect of specific features on the outcome.

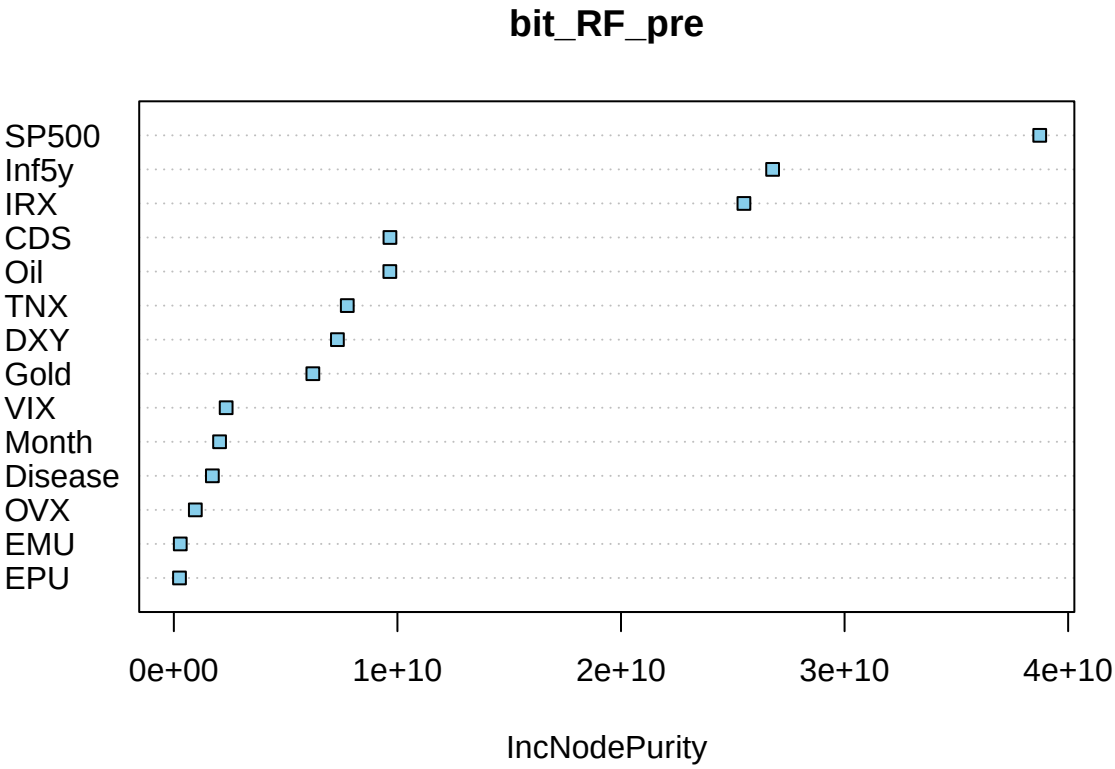


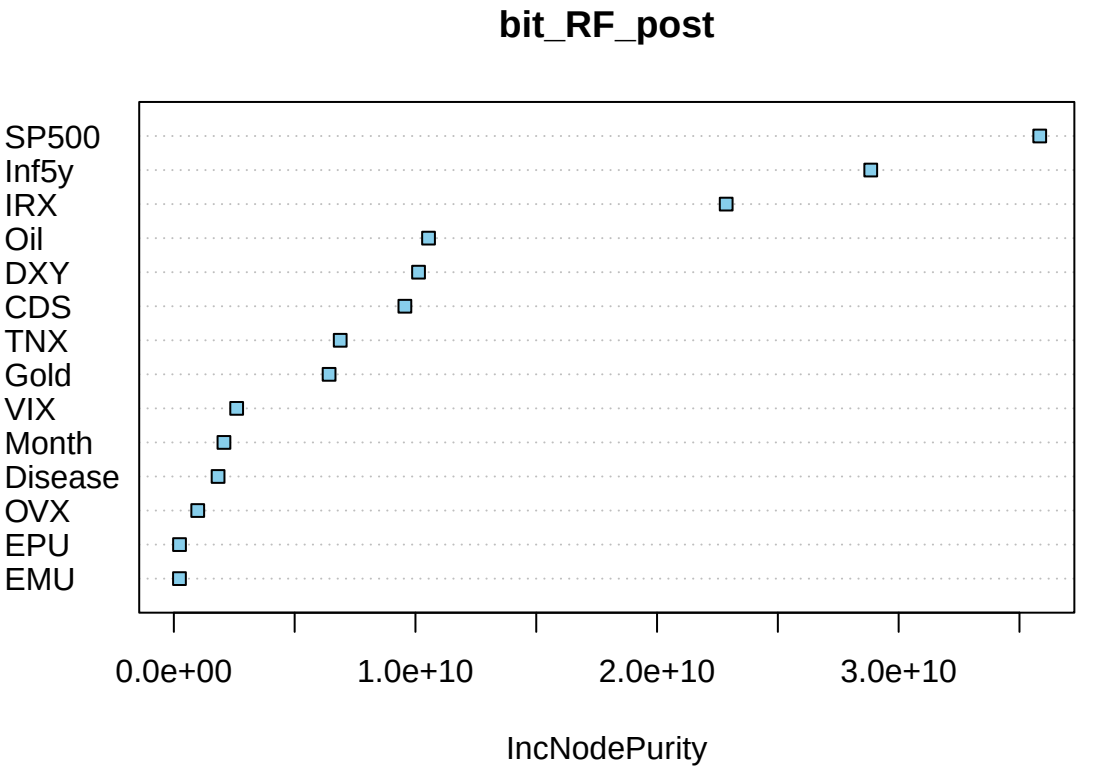
All variables, except for `IRX`, shows an increasing dependence plot. We can interpret `SP500`, `Inf5y`,`Gold`,`Disease` features have a positive effect on predicted outcome, and `IRX` has a negative effect on predicted outcome.

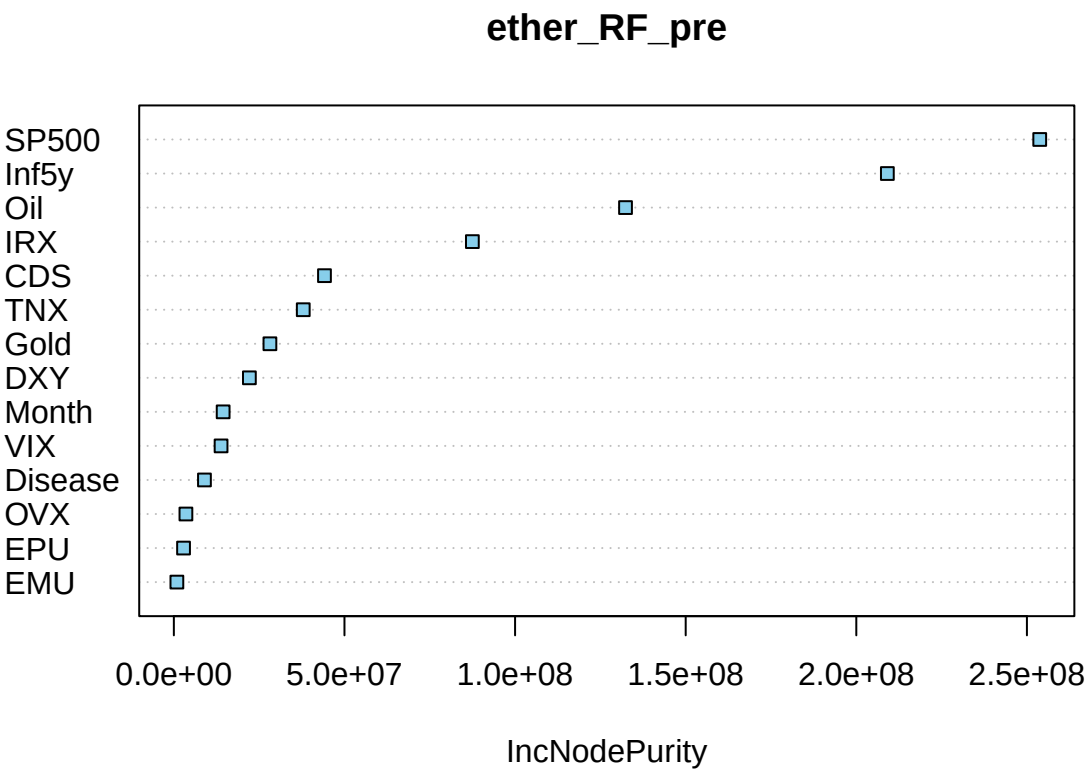
Table 2: Model performance with out-of-sample RMSEs (Ethereum)

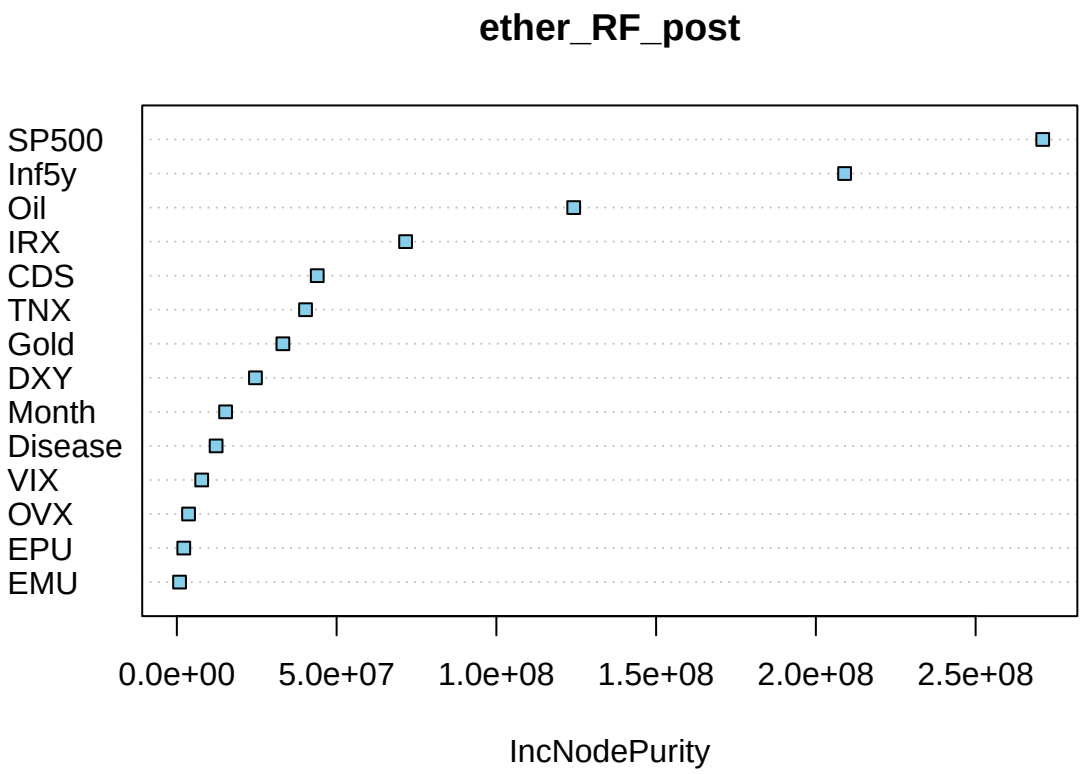
Model	RMSE
CART	2295.453
Random Forest	1076.176
Gradient Boosting	1283.753

Comparison for Pre-2019 and Post-2019






ether_RF_pre
SP500
Inf5y
Oil
IRX
CDS
TNX
Gold
DXY
Month
VIX
Disease
OVX
EPU
EMU
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
IncNodePurity



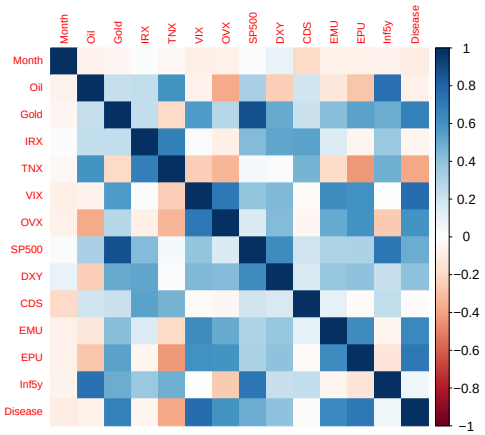
Conclusion

Appendix

Correlation Plots

For a deeper understanding of our paper, we will analyze the relationship with a correlation plot between predictors, which are macroeconomic and market performance factors. We know that correlation plot only measures the strength and direction of linear relationships between variables, but see correlation between predictors may provide some insights.

Below is the correlation plot for every factors we added, you can check there's some stronger correlations, represented with darker colors.



We decided to investigate the strong correlations with coefficients above 0.60.

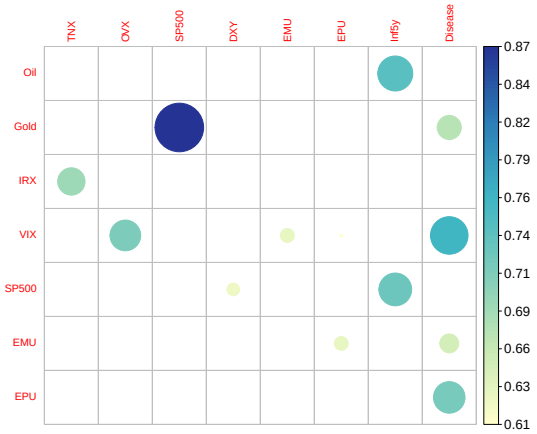


Table 3: Highest correlation among factors

Var1	Var2	Freq
Gold	SP500	0.8705935
VIX	Disease	0.7636956
Oil	Inf5y	0.7431183
SP500	Inf5y	0.7262370
EPU	Disease	0.7172680
VIX	OVX	0.7117934
IRX	TNX	0.6896120
Gold	Disease	0.6705888
EMU	Disease	0.6457534
VIX	EMU	0.6273703
EMU	EPU	0.6270346
SP500	DXY	0.6234803
VIX	EPU	0.6056454

Result shows `Gold` and `SP500` shows pretty strong correlation, 0.87. The next highest correlation is `VIX` and `Disease`.

(I think we should find some meaningful correlation here for our analysis)

	cluster1
SP500	4165.98043
Gold	1790.95124
EPU	126.51989
DXY	120.01837
EMU	97.30648

	cluster2
SP500	2454.34404
Gold	1263.72872
DXY	112.45689
EPU	84.70150
Oil	52.93657

	cluster3
SP500	2716.0546
Gold	1658.1686
EPU	489.8274
EMU	364.0126
OVX	148.6829

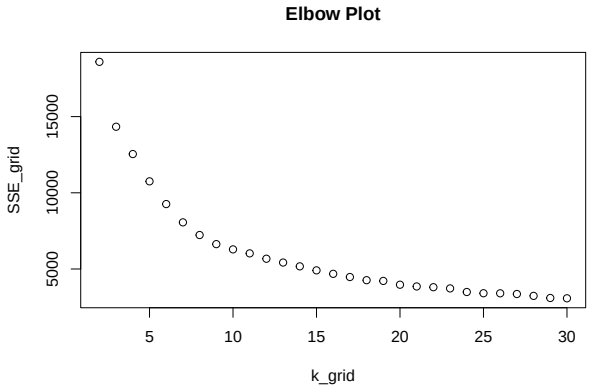
	cluster4
SP500	4016.2158
Gold	1897.4271
EMU	158.2080
EPU	153.9675
DXY	120.2058

	cluster5
SP500	3425.7795
Gold	1830.5749
EPU	253.4530
EMU	135.4127
DXY	115.6980

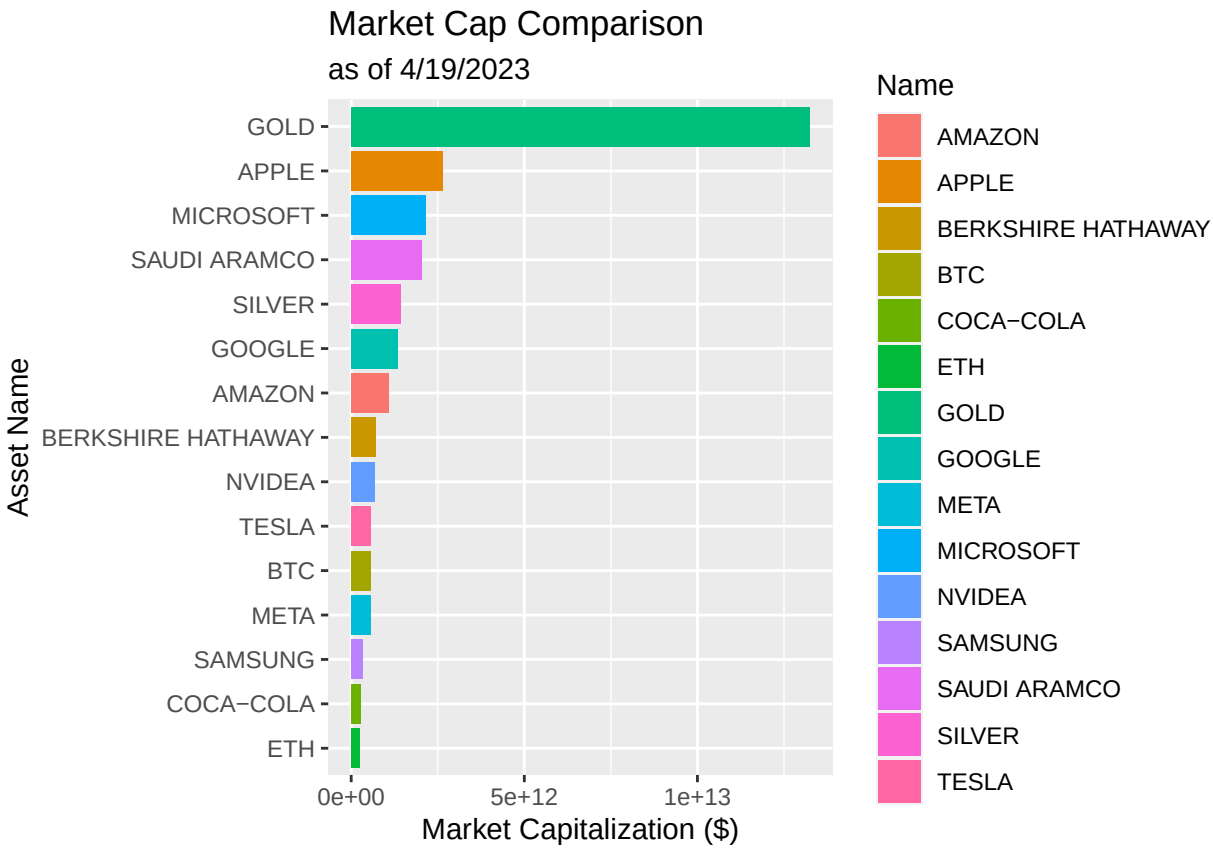
	cluster6
SP500	1926.08094
Gold	1268.64375
DXY	95.39114
Oil	94.16828
EPU	66.23844

K-means clustering

We did supervised learning method, Random Forest. Now we will try an unsupervised learning, K-means clustering method, which can be used to identify clusters of similar factors. First, will start from choosing optimal K, the amount of clusters. Below is Elbow plot. Elbow plot used to determine the optimal number of clustering. The plot displays within-cluster sum of squares(WSS) as a function of the number of clusters.



Will use 6 for k, the number of clusters.


Market Cap Comparison
as of 4/19/2023
Asset Name
GOLD
APPLE
MICROSOFT
SAUDI ARAMCO
SILVER
GOOGLE
AMAZON
BERKSHIRE HATHAWAY
NVIDEA
TESLA
BTC
META
SAMSUNG
COCA−COLA
ETH
0e+00
5e+12
1e+13
Market Capitalization ($)
Name
AMAZON
APPLE
BERKSHIRE HATHAWAY
BTC
COCA−COLA
ETH
GOLD
GOOGLE
META
MICROSOFT
NVIDEA
SAMSUNG
SAUDI ARAMCO
SILVER
TESLA