Critical Trading: Stakeholder Update

Team 2: Aditya Dhanotia, Tom Do, Jishnu Channamsetti, Shantanu Trivikram, Levon Hakobyan, Zhixing Zhu, Colin Zhou

Agenda

Experiment

- Objectives
- Hypothesis

Data Information

- Shape of Data

EDA

- Summary Statistics
- Correlations

Data Cleaning

- Winsorization

Feature Engineering

- PCA

Model Selection

- Linear
- Non-Linear
- Deep Learning

What is QQQ?

- PowerShares QQQ ETF
- The 100 largest non-financial companies listed on the NASDAQ stock exchange.
- Popular for its exposure to tech giants like Apple, Microsoft, Amazon, etc.
- Often used by investors to gain exposure to tech sector and large-cap growth stocks.

Nasdaq 100 QQQ Components

#	Company	Symbol	Weight	Price	Chg	% Chg
1	Apple Inc	AAPL	11.041	172.88	-2.57	(-1.46%)
2	Microsoft Corp	MSFT	9.914	326.67	-4.65	(-1.40%)
3	Amazon.com Inc	AMZN	5.276	125.17	-3.23	(-2.52%)
4	NVIDIA Corp	NVDA	4.176	413.87	-7.14	(-1.70%)
5	Meta Platforms Inc	META	3.928	308.65	-4.16	(-1.33%)
6	Alphabet Inc	GOOGL	3.286	135.60	-2.15	(-1.56%)
7	Alphabet Inc	GOOG	3.24	136.74	-2.21	(-1.59%)
8	Broadcom Inc	AVGO	3.086	853.63	-14.20	(-1.64%)
9	Tesla Inc	TSLA	2.749	211.99	-8.12	(-3.69%)
10	Adobe Inc	ADBE	2.16	540.96	-14.78	(-2.66%)
11	Costco Wholesale Corp	COST	2.148	552.93	-12.70	(-2.25%)
12	PepsiCo Inc	PEP	1.93	160.00	-0.56	(-0.35%)
13	Cisco Systems Inc	CSCO	1.89	52.93	-0.39	(-0.73%)
14	Netflix Inc	NFLX	1.557	400.96	-0.81	(-0.20%)
15	Comcast Corp	CMCSA	1.545	42.86	-0.21	(-0.49%)
16	Advanced Micro Devices Inc	AMD	1.441	101.81	-0.59	(-0.58%)
17	T-Mobile US Inc	TMUS	1.412	136.99	-0.85	(-0.62%)
18	Amgen Inc	AMGN	1.307	278.81	-1.79	(-0.64%)
19	Intel Corp	INTC	1.281	34.92	-0.75	(-2.10%)
20	Intuit Inc	INTU	1.243	506.81	-14.71	(-2.82%)

Experiment Overview

Objective:

- Assess the predictability of the realized volatility of the QQQ ETF using a set of macro features.
- This experiment aims to build a robust model that can assist in making informed investment decisions.

Hypothesis:

- Macro features such as GDP, equities, credit spread, futures, implied volatility, and returns can provide significant predictive power for the realized volatility of QQQ.

About the Data

783
Total predictors

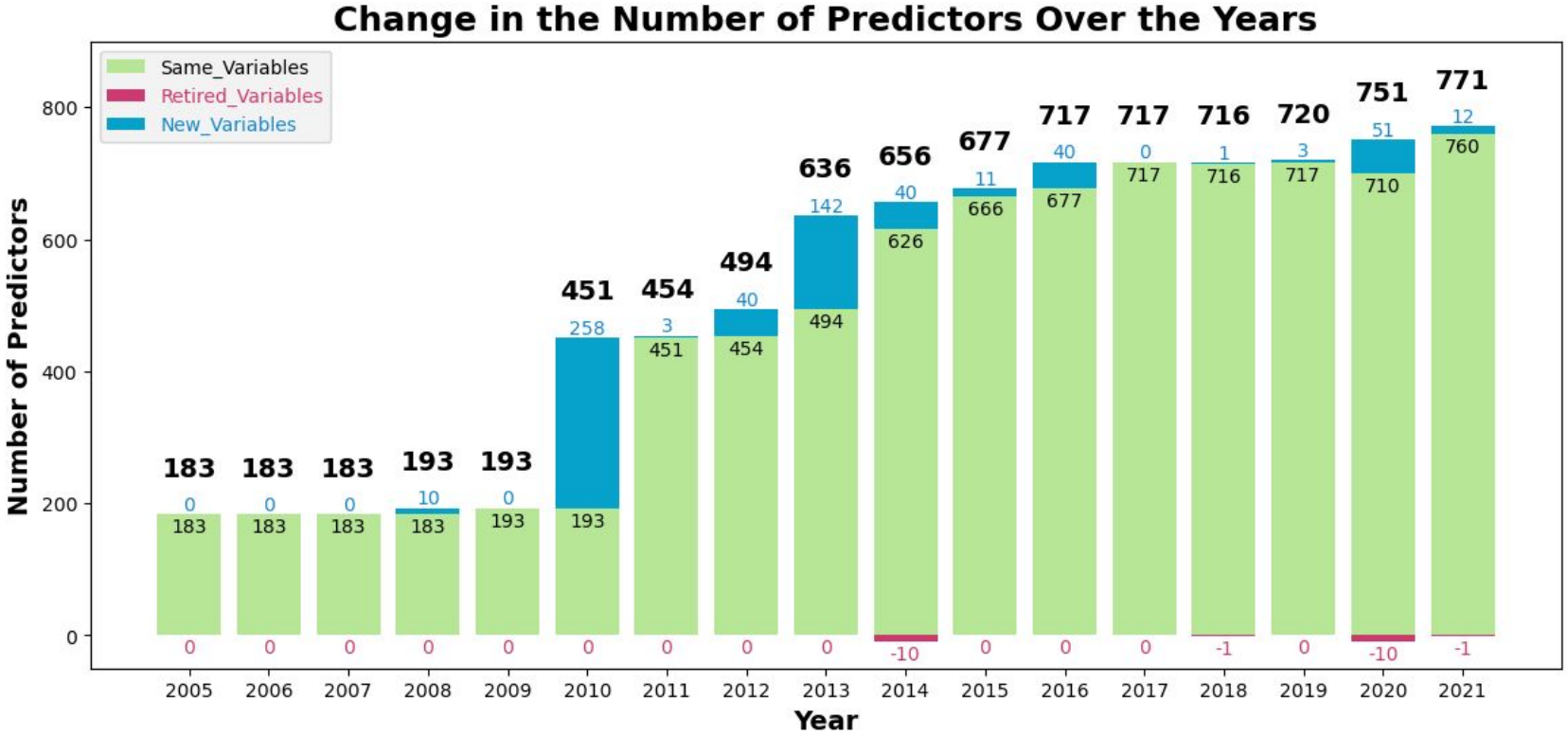
4,160
Recorded trading dates
(2005-2021)

To predict
Realized Volatility
22 Days Later

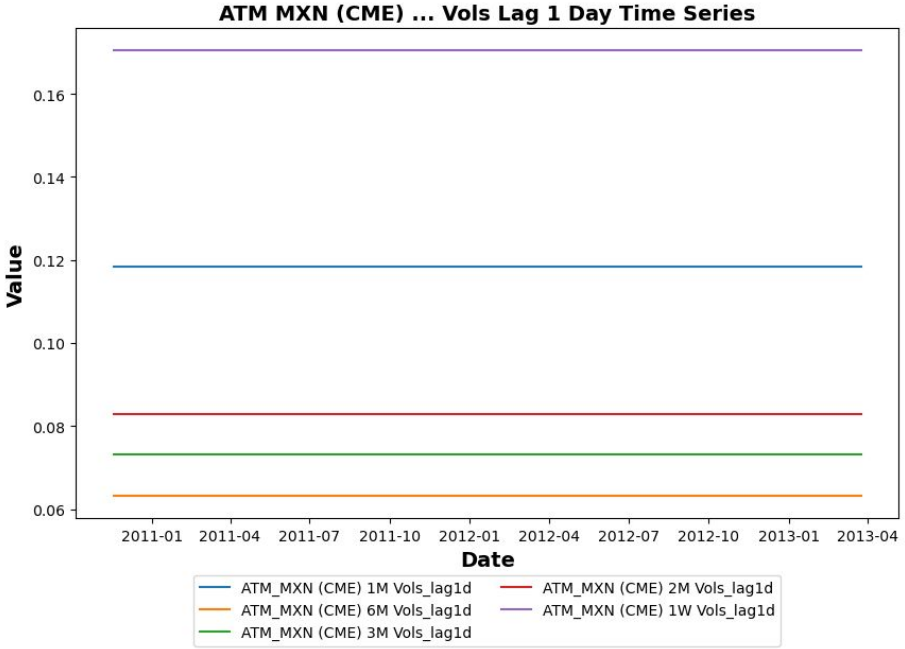
182
Original predictors
surviving since 2005

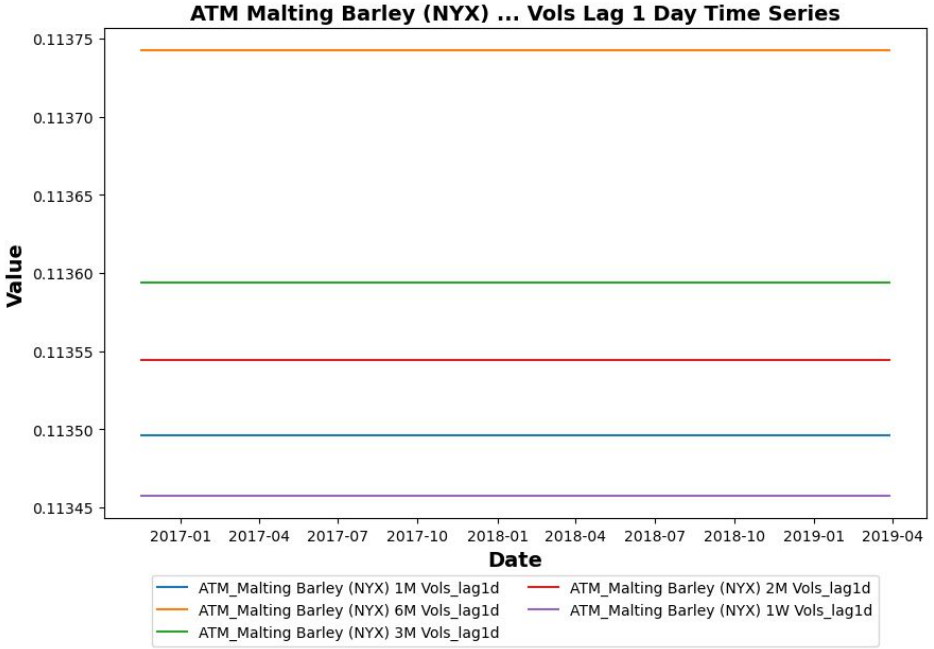
881
Rolling dates used to
build/test daily models

Predictors Evolved Over Time



Why Were Predictors Dropped?

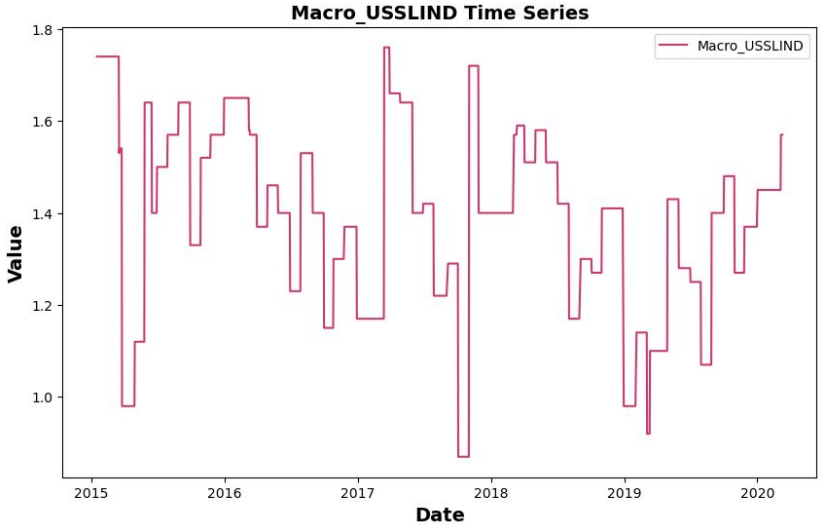




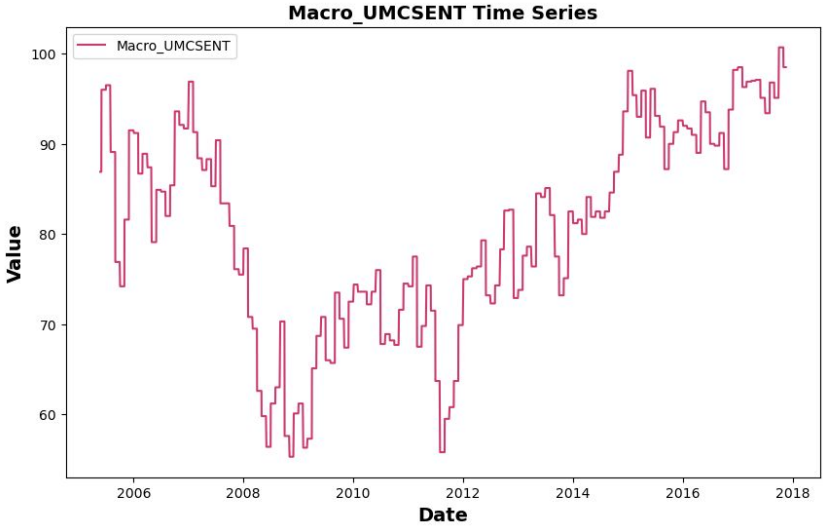
They were **constant (NAN correlation)**

Why Were 12 Predictors Dropped?

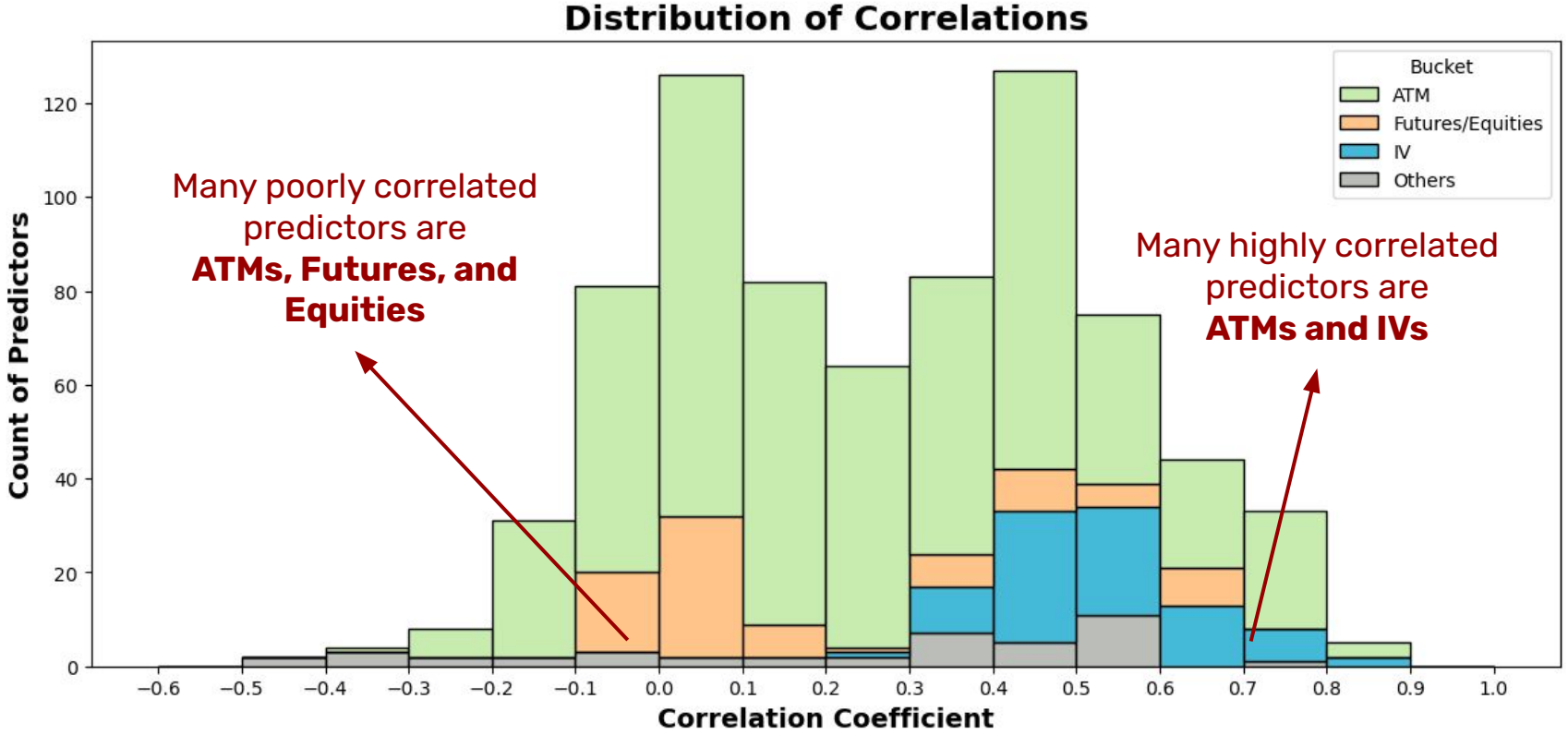
Leading Index for the United States



University of Michigan Consumer Sentiment

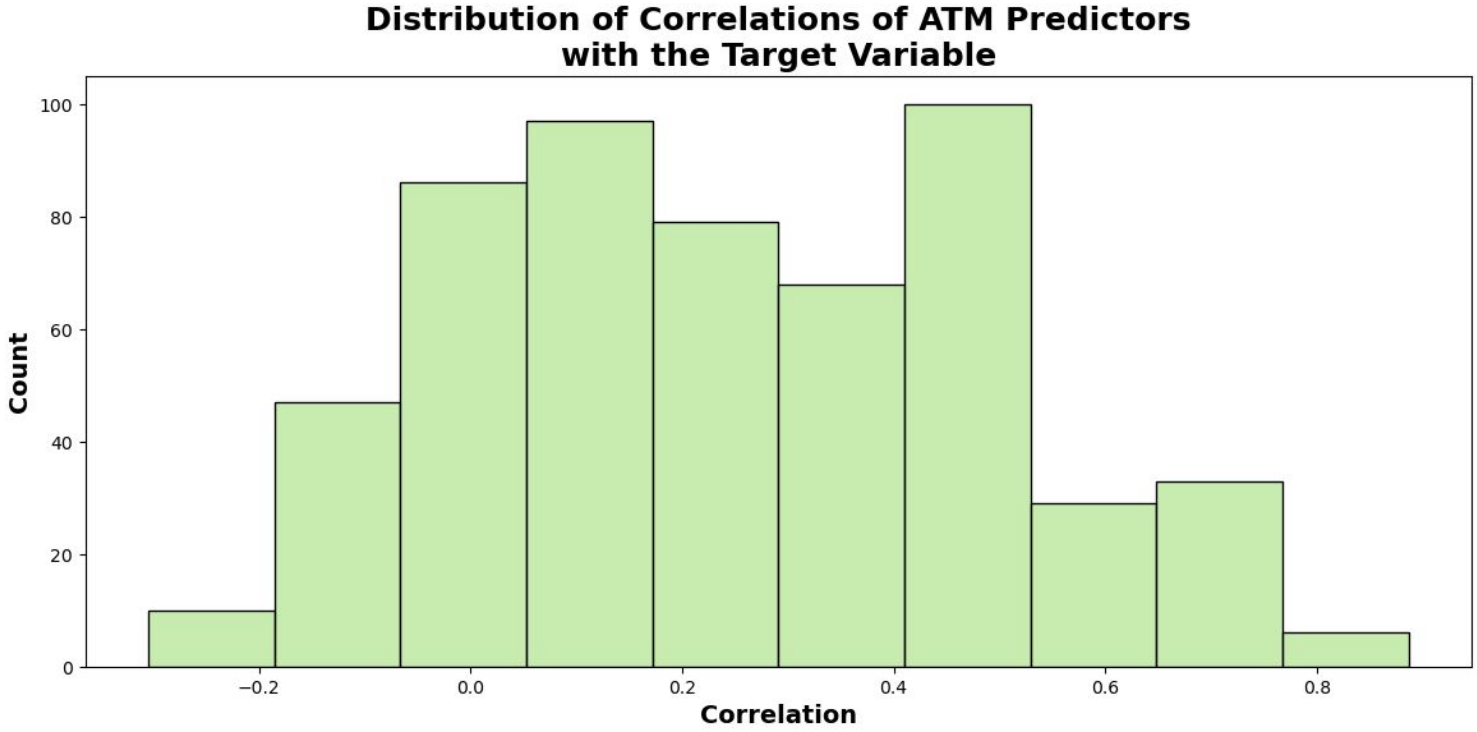


Most Positive Correlation: IV and ATM



ATM Predictors

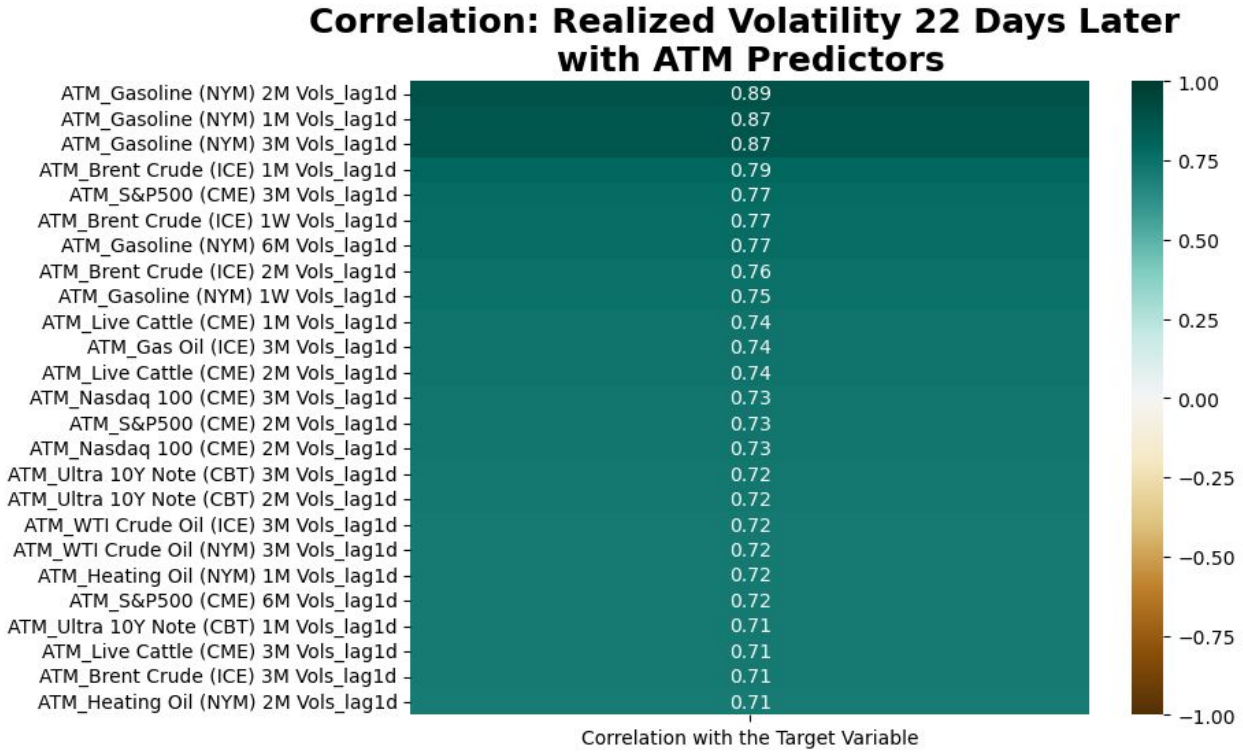
560
out of 783
predictors



ATM Energy Predictors Have High Impact

51

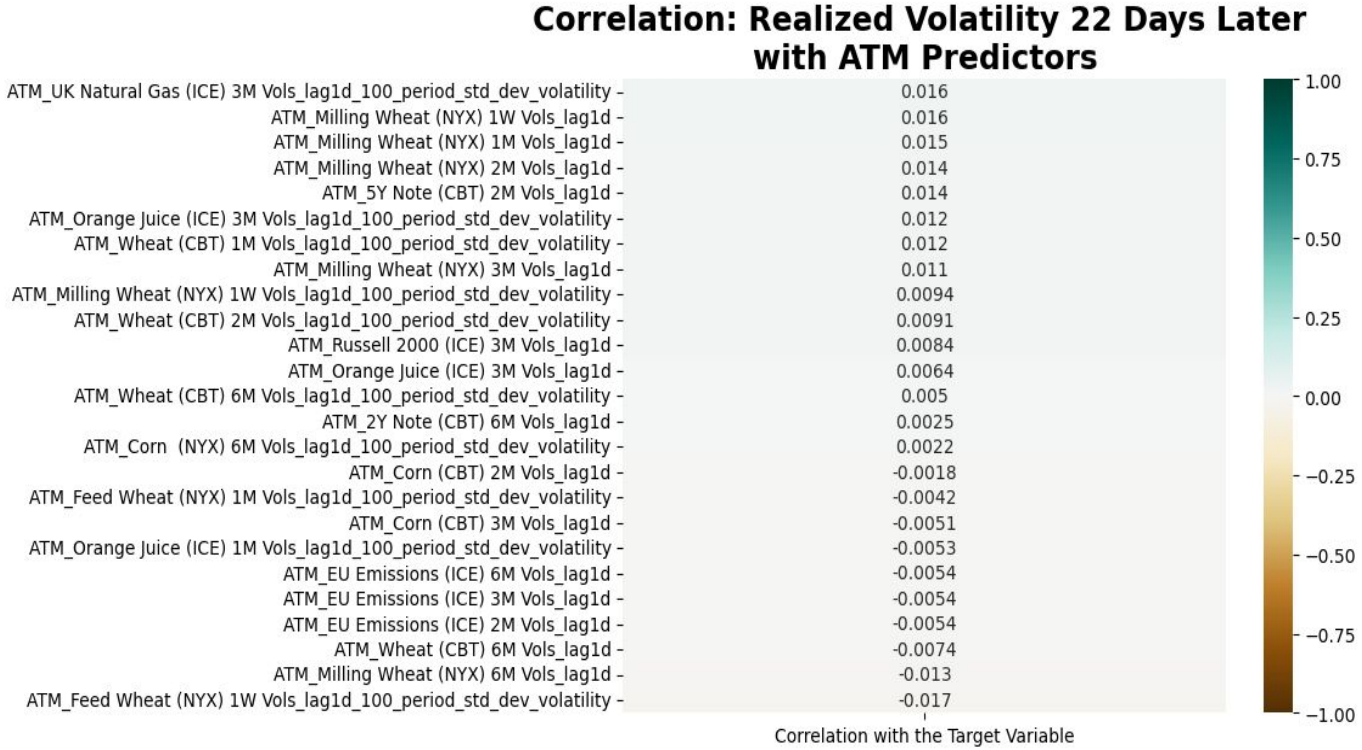
out of 560 ATM Predictors has an absolute correlation more than 0.6



ATM Agri/Emission Predictors Have Less Impact

155

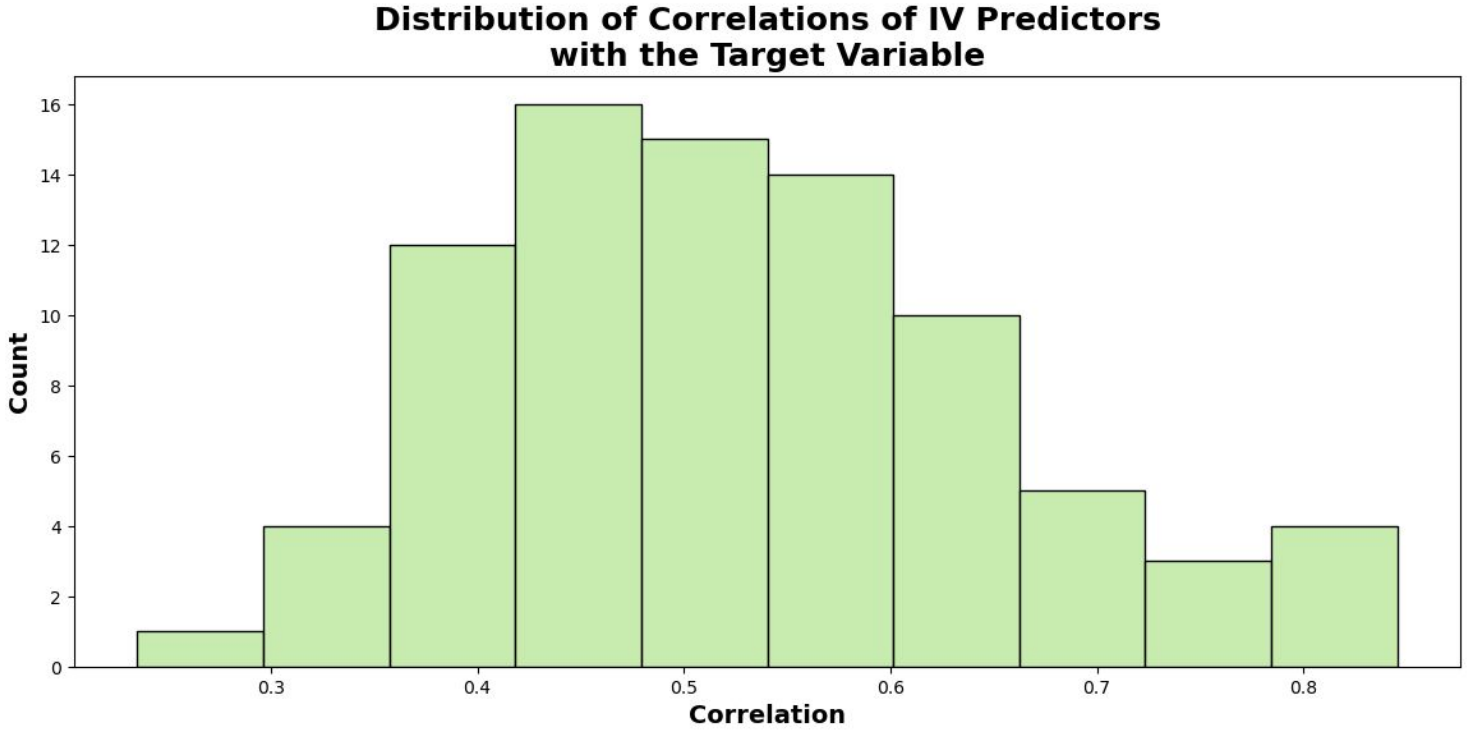
out of 560 ATM Predictors has an absolute correlation less than 0.1



IV Predictors: Highly Correlated

84

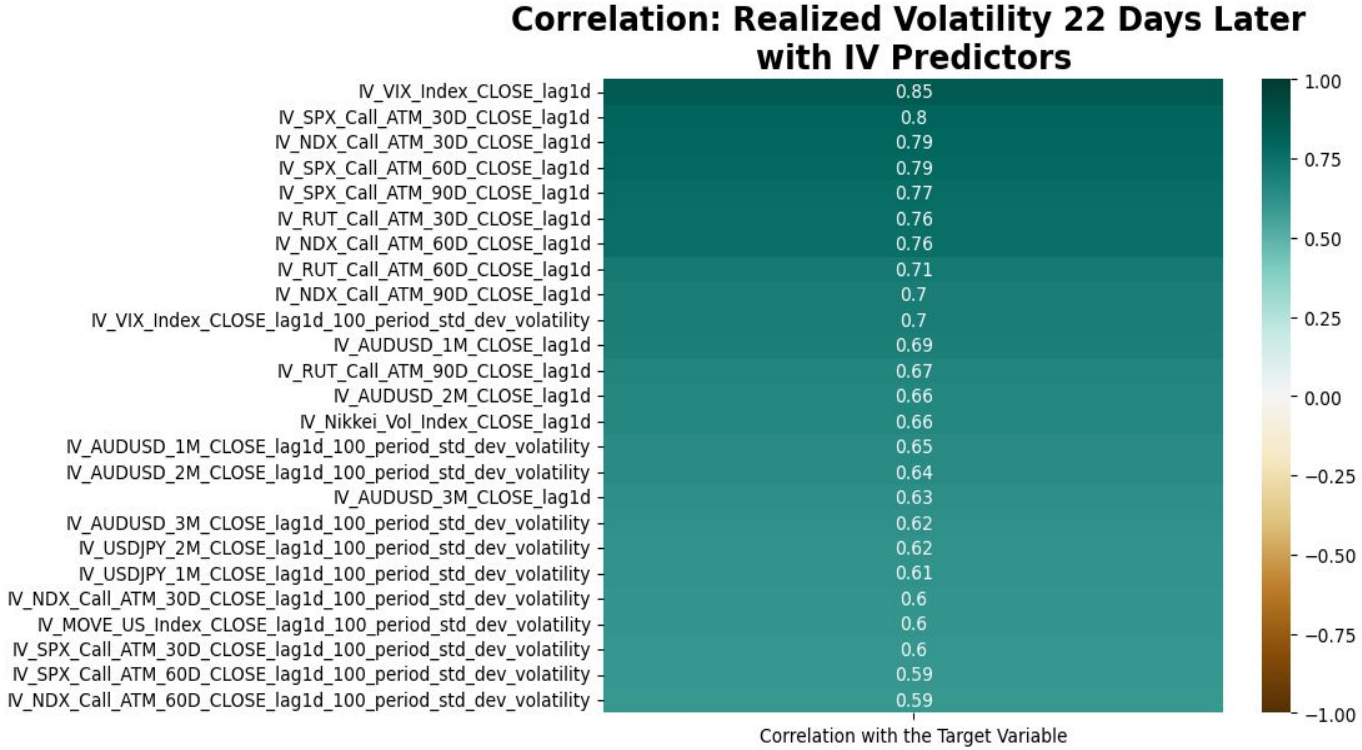
out of 783 predictors



IV Predictors: Highly Correlated

22

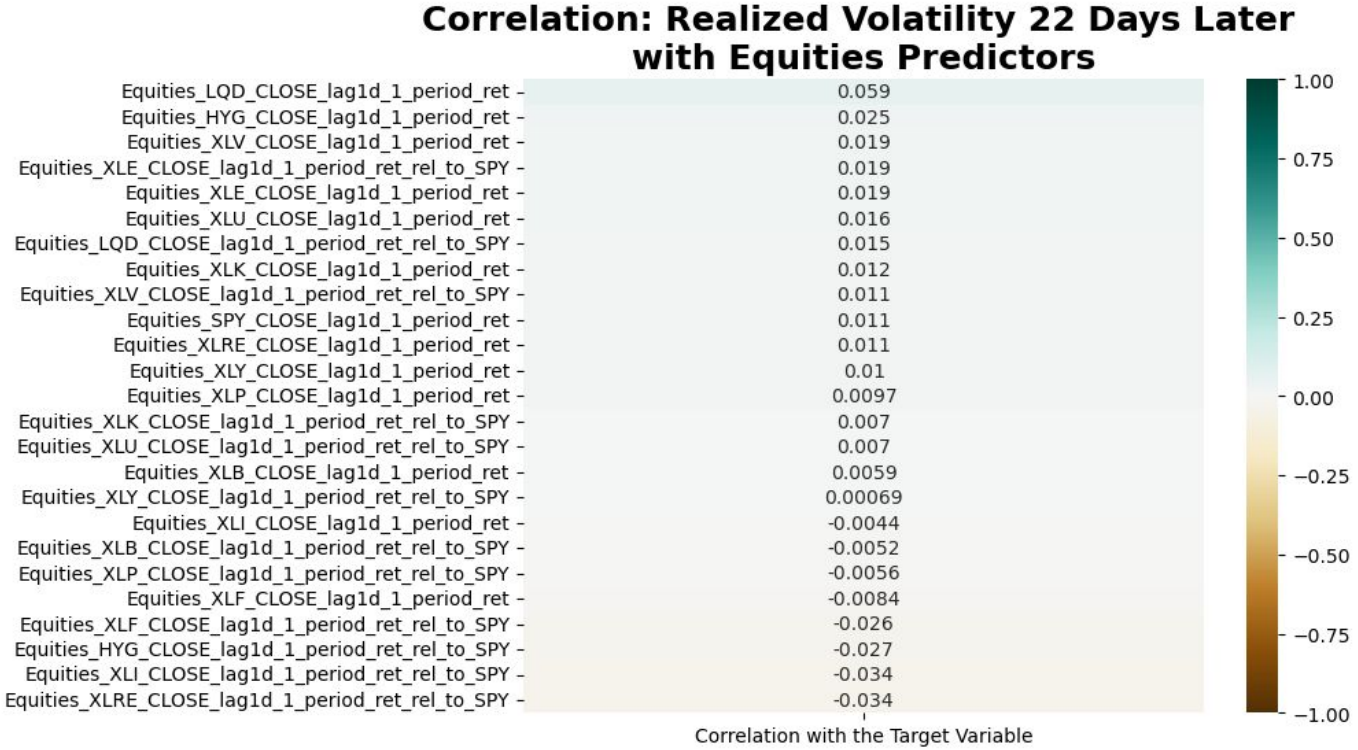
out of 84
IV Predictors
has an absolute
correlation
more than 0.6



Equities Predictors: Many Poorly Correlated

25

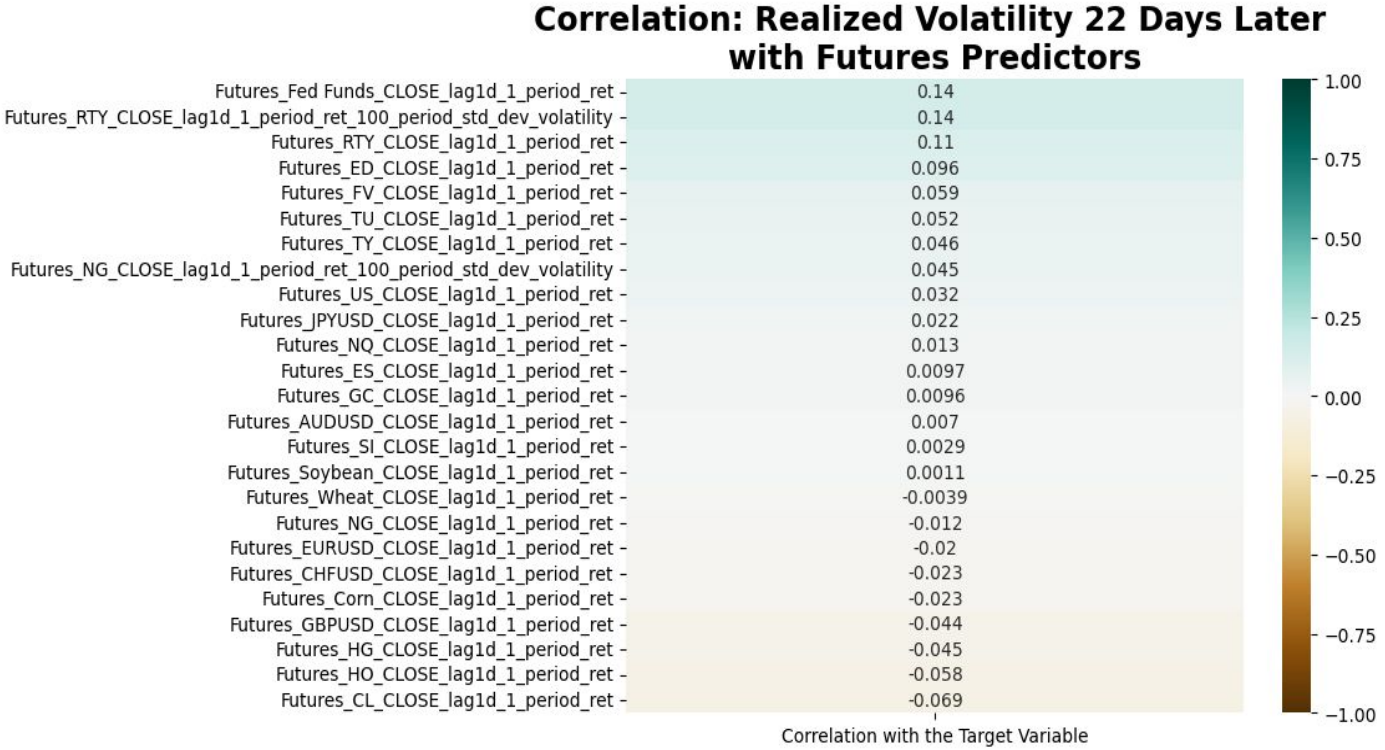
out of 38 Equities Predictors has an absolute correlation less than 0.1



Futures Predictors: Many Poorly Correlated

22

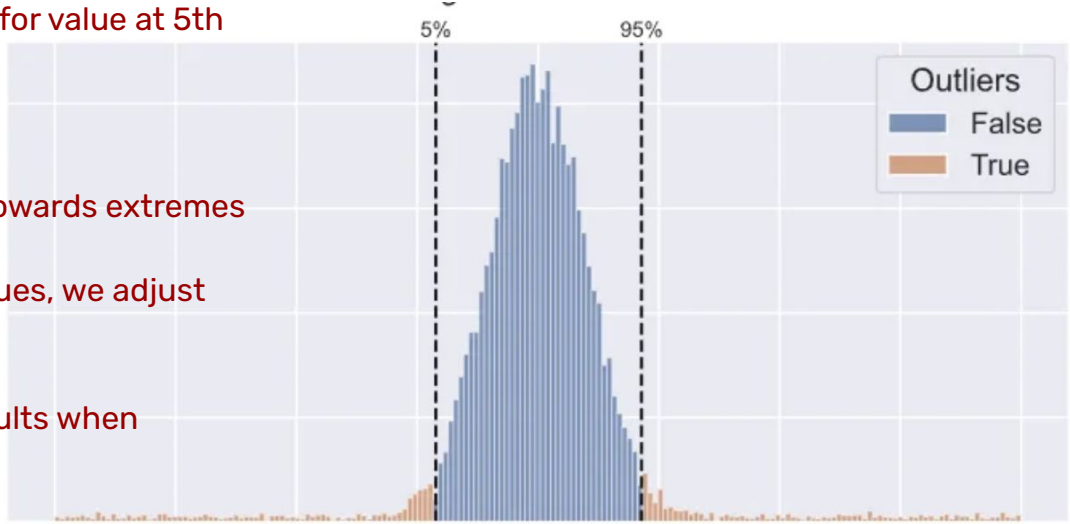
out of 46 Futures Predictors has an absolute correlation less than 0.1



Outlier Detection
Why Winsorization?

Investigating outliers and bad data

- Instead of removing extreme values that distort the overall picture of data, we adjust them
- 16 variables have max value greater than 1^10
- 95% of distribution is taken - high extreme values are changed for the value at 95th percentile, low values changed for value at 5th percentile



Benefits of Winsorization:

Reduces Skewness: Makes our data less tilted towards extremes

Preserves Data: Instead of deleting extreme values, we adjust them, so we don't lose information

Robust Analysis: Helps in producing reliable results when analyzing data, especially in statistical analysis

Findings after Winsorization

After adjusting for outliers following variables persisted with still very high extreme values greater than 1^10:

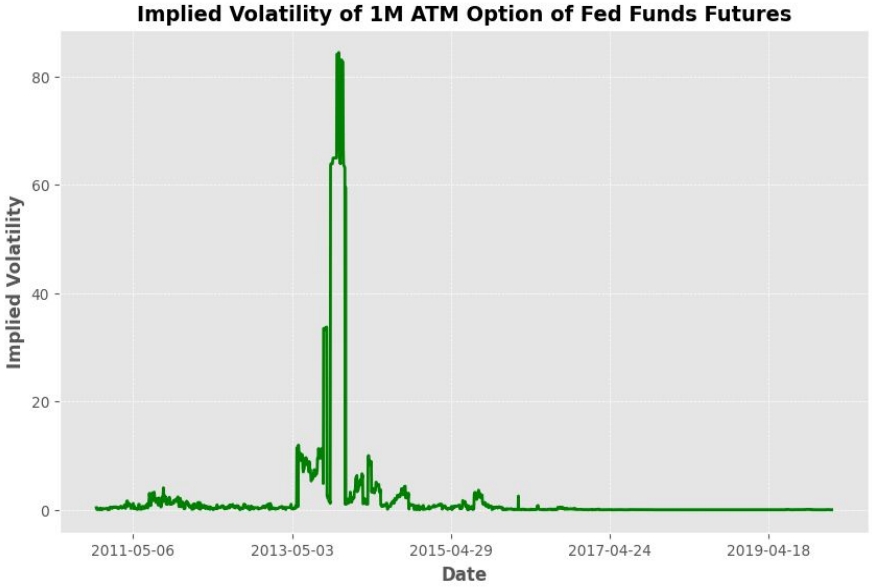
1. ***Implied Volatility of (1, 2, 3 and 6 month) ATM option on Fed Funds Futures***

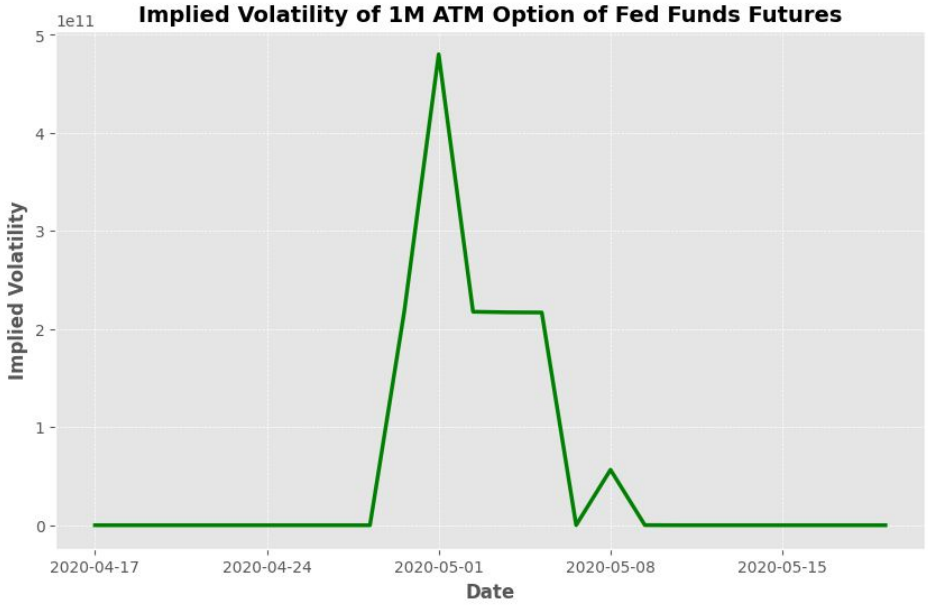
1. ***Standard Deviation of Implied Volatility over 100 days of (1,2,3, and 6 month) Fed Funds Futures***

1. ***Implied Volatility of (1, 2, 3 and 6 month) ATM option on EU Emission allowances***

1. ***Standard Deviation of Implied Volatility over 100 days of ATM EU Emission allowances***

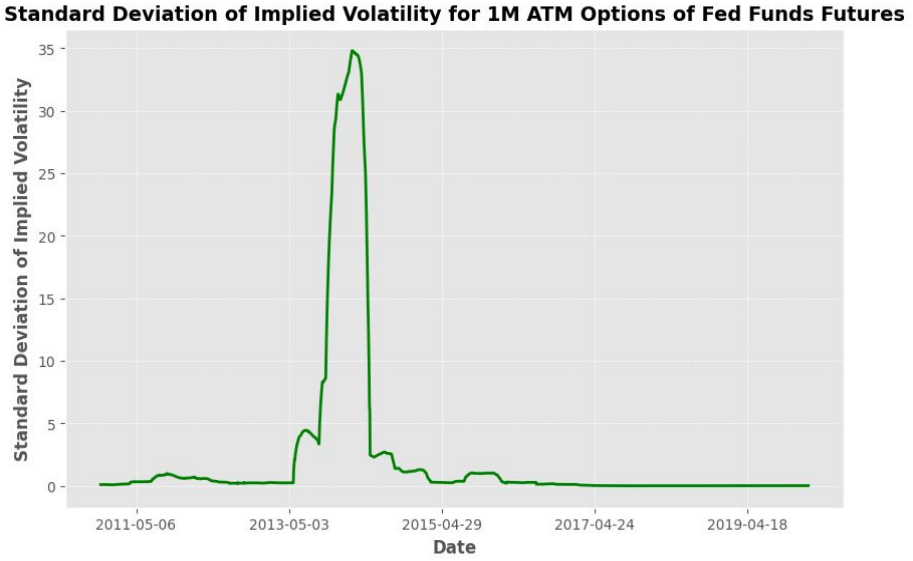
Analyzing Extreme Values

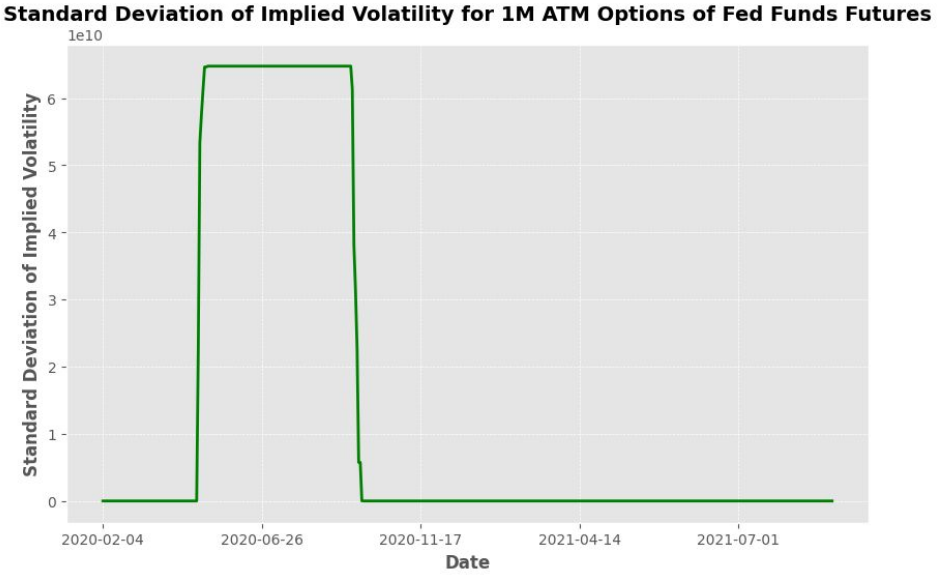




IV of 1M ATM Option of Fed Funds before and after 2020

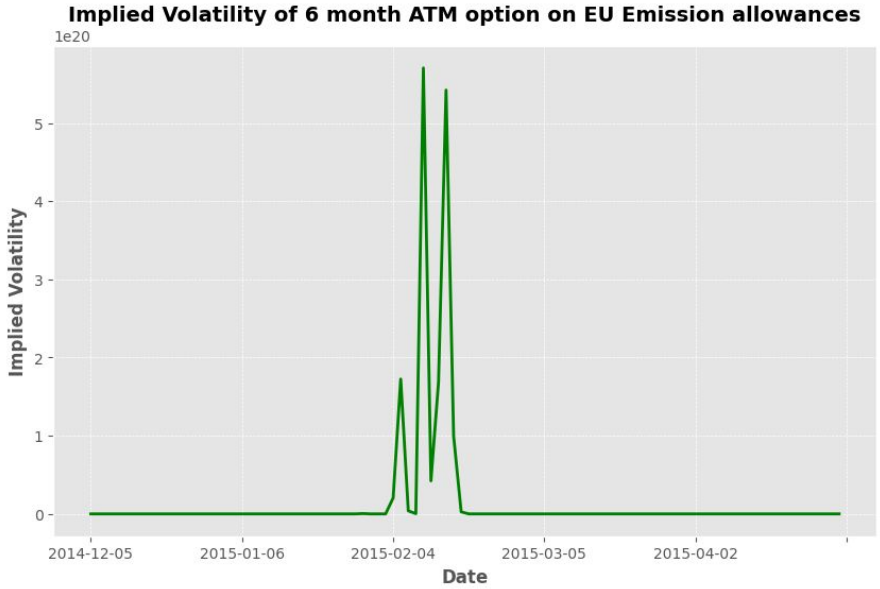
Analyzing Extreme Values

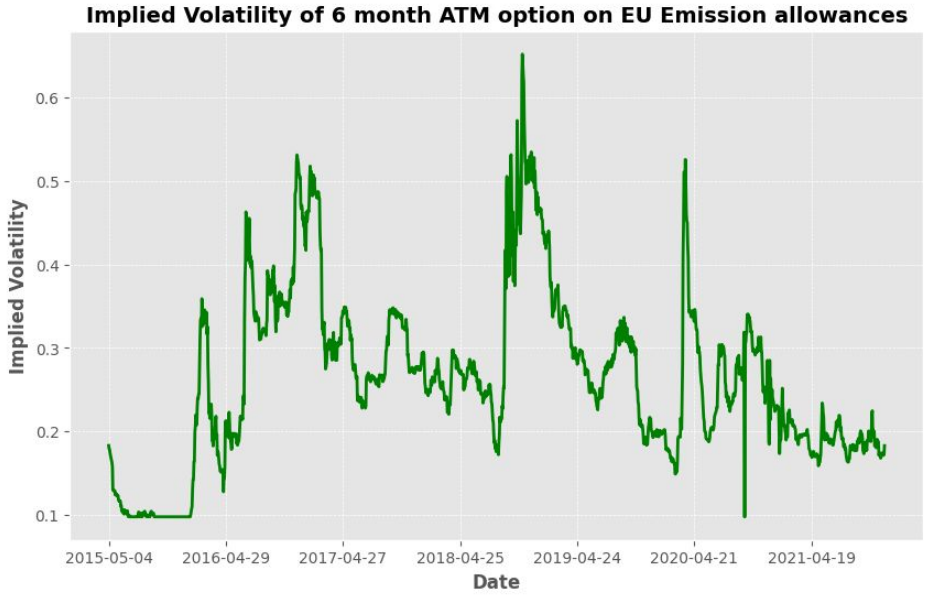




Standard Deviation of IV for 1 M ATM Options of Fed Funds Futures

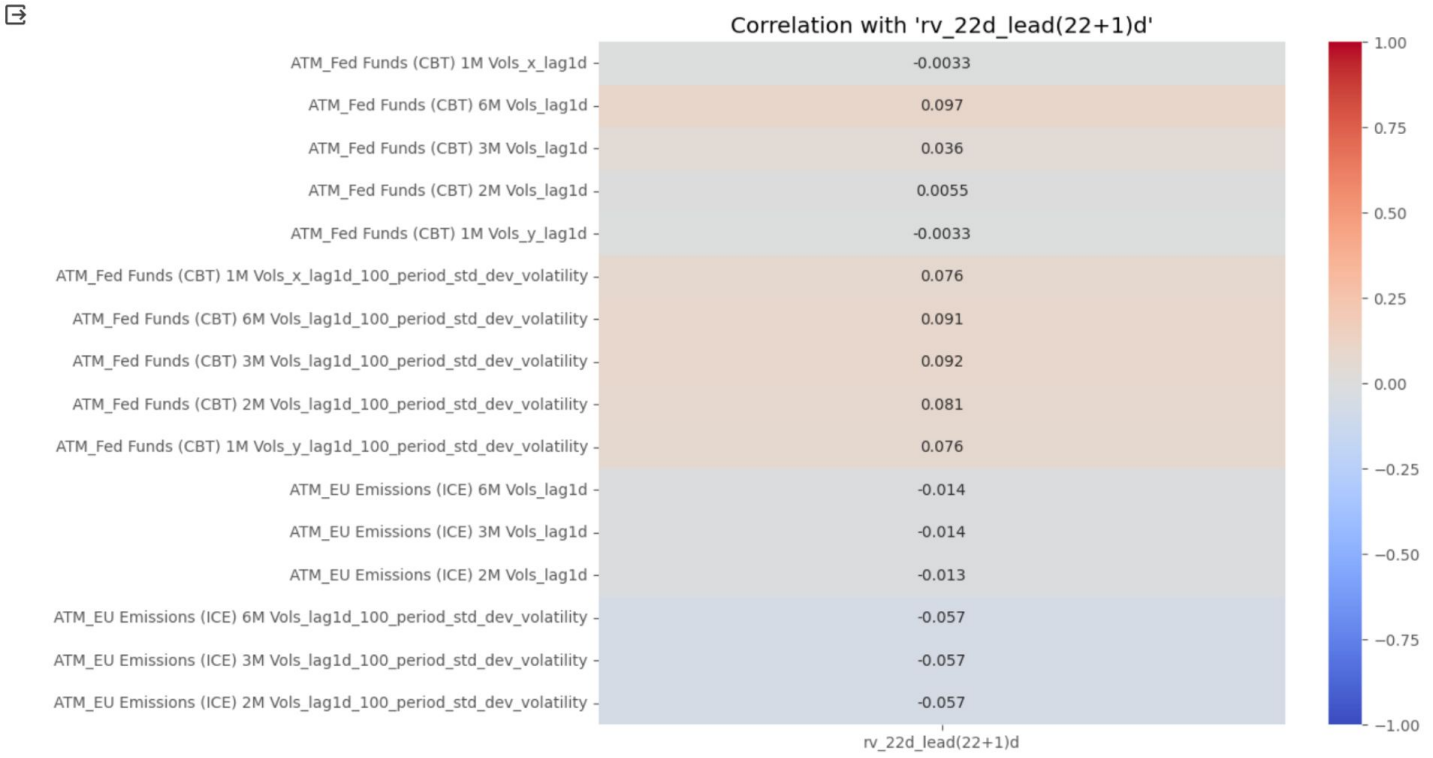
Analyzing Extreme Values





Implied Volatility of 6M ATM option on EU Emission Allowance before and after 2015

Analyzing Extreme Values



Summary Stats

Top 5 by Mean

	count	mean	std	skewness	kurtosis	median	cv
Macro_GDP	4160.0	16951.764053	2922.447097	0.350675	-1.155825	16661.000000	0.172398
IV_MOVE_US_Index_CLOSE_lag1d	4160.0	7890.131354	8077.690643	2.824779	8.330530	5099.545326	1.023771
IV_2Y_Swaps_1M_CLOSE_lag1d	4160.0	3278.572094	2709.357087	0.784154	0.045125	3012.363250	0.826383
IV_2Y_Swaps_3M_CLOSE_lag1d	4160.0	3213.980037	2805.118463	1.468974	3.112642	3056.431250	0.872787
IV_2Y_Swaps_6M_CLOSE_lag1d	4160.0	3029.938440	2482.290619	1.291028	2.546951	3021.152450	0.819254

Top 5 by SD

	count	mean	std	skewness	kurtosis	median	cv
IV_MOVE_US_Index_CLOSE_lag1d	4160.0	7890.131354	8077.690643	2.824779	8.330530	5099.545326	1.023771
IV_MOVE_US_Index_CLOSE_lag1d_100_period_std_dev_volatility	4160.0	2236.492847	3100.897712	3.053581	9.474922	1105.452210	1.386500
IV_30Y_Swaps_1M_CLOSE_lag1d_100_period_std_dev_volatility	4160.0	894.025065	2922.694842	5.241530	26.856119	187.353022	3.269142
Macro_GDP	4160.0	16951.764053	2922.447097	0.350675	-1.155825	16661.000000	0.172398
IV_2Y_Swaps_3M_CLOSE_lag1d	4160.0	3213.980037	2805.118463	1.468974	3.112642	3056.431250	0.872787

Summary Stats

Top 5 by skewness

	count	mean	std	skewness	kurtosis	median	cv
ATM_AUD (CME) 1W Vols_lag1d	2771.0	0.015066	0.107622	52.008532	2726.547863	0.010308	7.143401
ATM_Soybean Meal (CBT) 1W Vols_lag1d	2771.0	0.067368	0.239009	49.488821	2544.946579	0.054859	3.547825
ATM_WTI Crude (NYM) 1W Vols_lag1d	2157.0	0.225329	1.404972	31.867281	1216.956107	0.096314	6.235191
ATM_WTI Crude Oil (NYM) 1M Vols_lag1d	2157.0	0.179338	0.827515	30.446218	1139.669455	0.096648	4.614266
ATM_WTI Crude (ICE) 1W Vols_lag1d	2314.0	0.215012	1.239153	30.219120	1126.207663	0.093539	5.763178

Top 5 by Kurtosis

	count	mean	std	skewness	kurtosis	median	cv	
ATM_AUD (CME) 1W Vols_lag1d	2771.0	0.015066	0.107622	52.008532	2726.547863	0.010308	7.143401	(
ATM_Soybean Meal (CBT) 1W Vols_lag1d	2771.0	0.067368	0.239009	49.488821	2544.946579	0.054859	3.547825	(
ATM_WTI Crude (NYM) 1W Vols_lag1d	2157.0	0.225329	1.404972	31.867281	1216.956107	0.096314	6.235191	(
ATM_WTI Crude Oil (NYM) 1M Vols_lag1d	2157.0	0.179338	0.827515	30.446218	1139.669455	0.096648	4.614266	(
ATM_WTI Crude (ICE) 1W Vols_lag1d	2314.0	0.215012	1.239153	30.219120	1126.207663	0.093539	5.763178	(

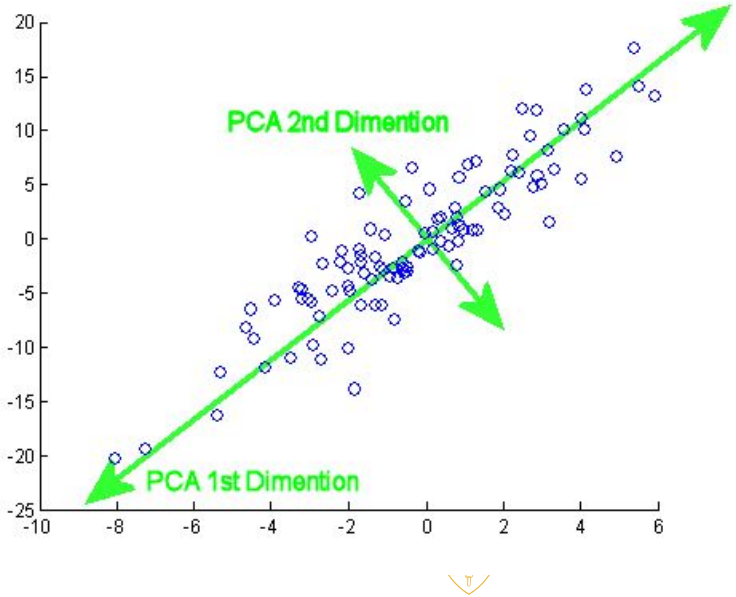
Summary Stats

Top 5 by CV

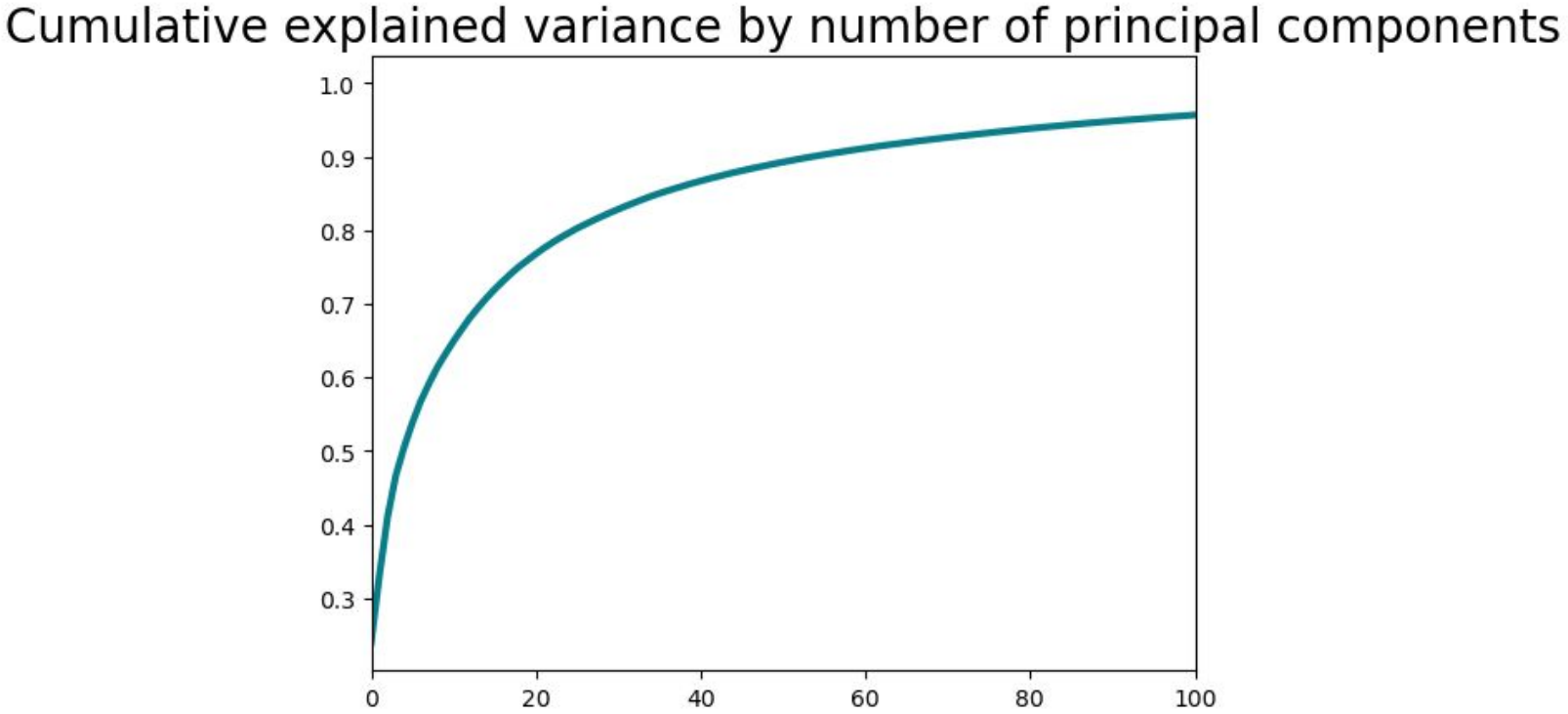
	count	mean	std	skewness	kurtosis	median	cv	
Futures_NG_CLOSE_lag1d_1_period_ret	4160.0	1.182095e-04	0.028412	0.178429	0.010794	-0.000375	240.353156	-
Futures_AUDUSD_CLOSE_lag1d_1_period_ret	4160.0	5.801422e-05	0.006787	-0.123231	0.120237	0.000251	116.986615	-
Futures_Fed Funds_CLOSE_lag1d_1_period_ret	4160.0	4.008043e-07	0.000043	-0.029496	5.916947	0.000000	107.665139	-
Futures_CHFUSD_CLOSE_lag1d_1_period_ret	4160.0	6.242132e-05	0.005415	0.059964	-0.055952	-0.000042	86.749143	-
Futures_CL_CLOSE_lag1d_1_period_ret	4160.0	3.008296e-04	0.020724	-0.085281	0.159840	0.000905	68.891011	-

Principal Component Analysis

- Dimensionality Reduction
- Explains Maximum Variance
- Does not lose key characteristics
- Less Complexity



Principal Component Analysis



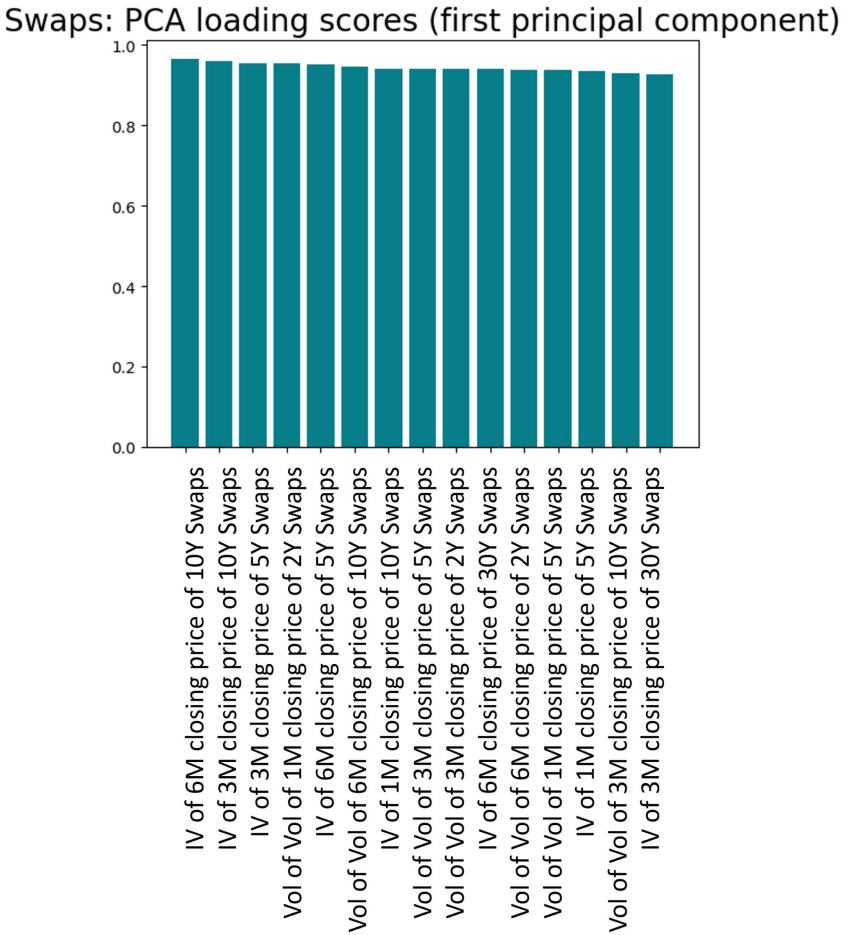
Bucketing

- Equities
- Futures
- Commodities
- Swaps
- Currencies
- Additional (Returns, Standard Deviation, IV, Volatility of IV, ATM)

Swaps

Implied Volatility and Vol of Vol are the most defining variables for explaining Swaps

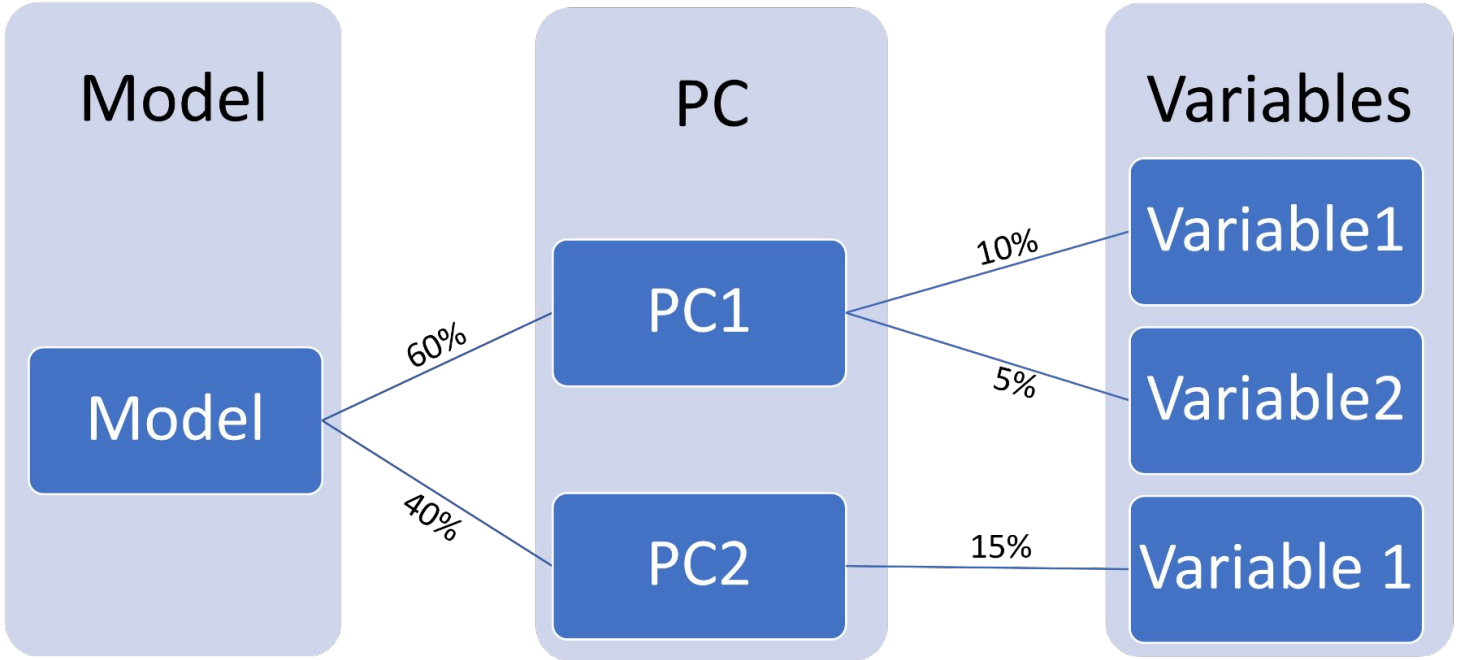
IV of 3M and 6M Close of 10Y Swaps most conclusive



Most Contributing Variables

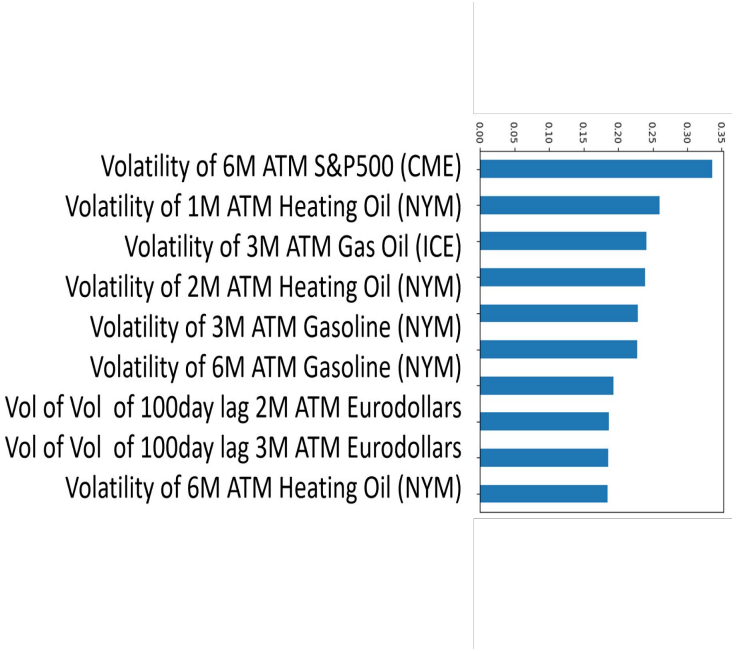
Buckets	Variable Name
Equities	Std Deviation over 100 day period for returns prices of SPY Close
Equities	Std Deviation over 100 day period for returns prices of XLI Close
Futures	Std Deviation over 100 day period for return prices of AUD-USD Close
Futures	Std Deviation over 100 day period for return prices of EUR-USD Close
Commodities	Std Deviation over 100 day period for volatility of 3M ATM Gas Oil (ICE)
Commodities	Std Deviation over 100 day period for volatility of 3M ATM Heating Oil (NYM)
Currency	Implied Volatility of 3M EUR-USD Close
Currency	Implied Volatility of 2M EUR-USD Close

Weighted Average of Contribution of Variables

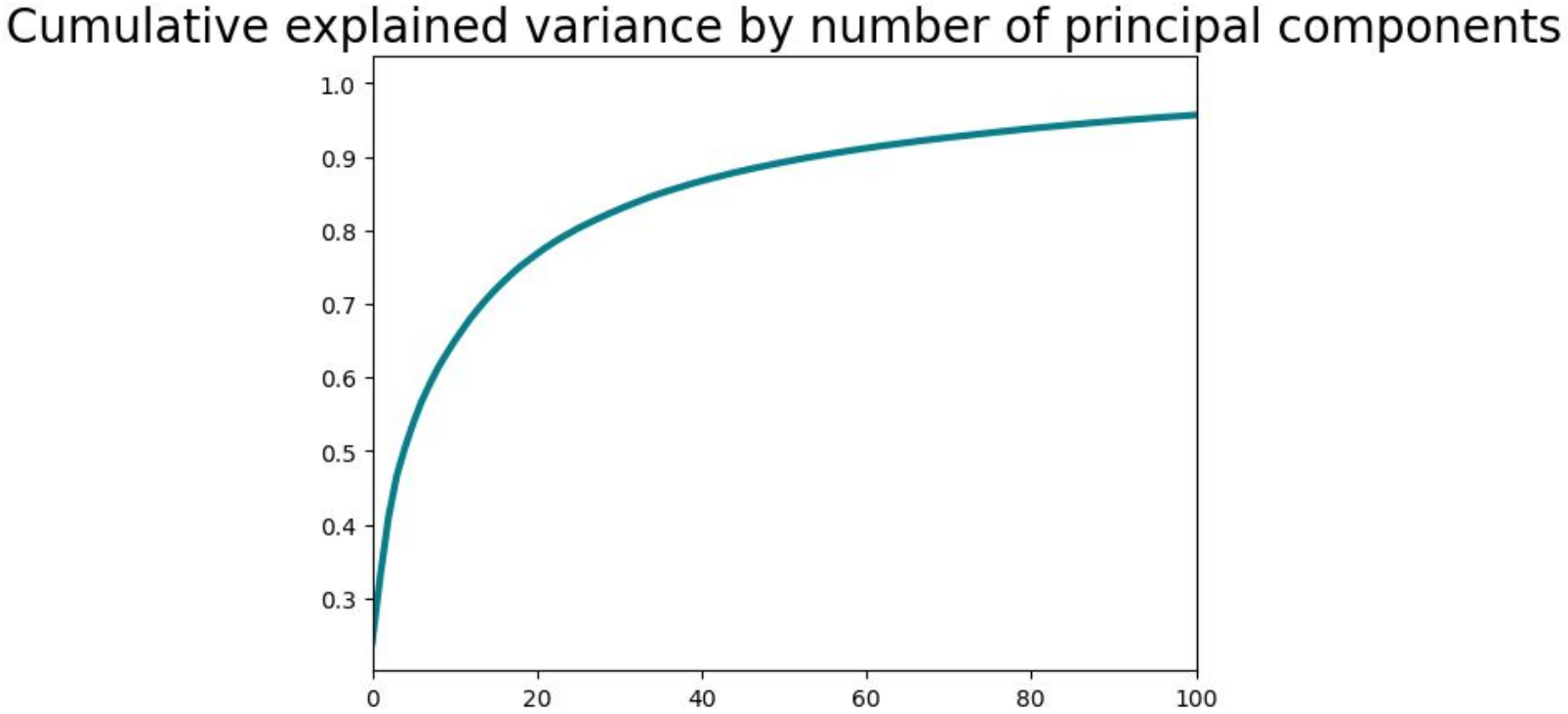


Variables contributing the most to all PCs (Weighted Average)

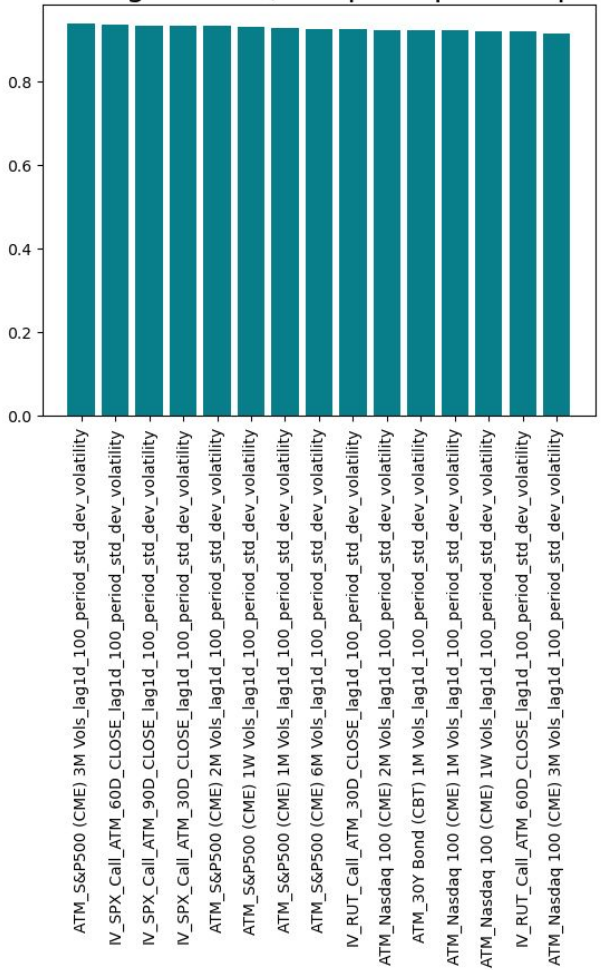
Correlation to Target



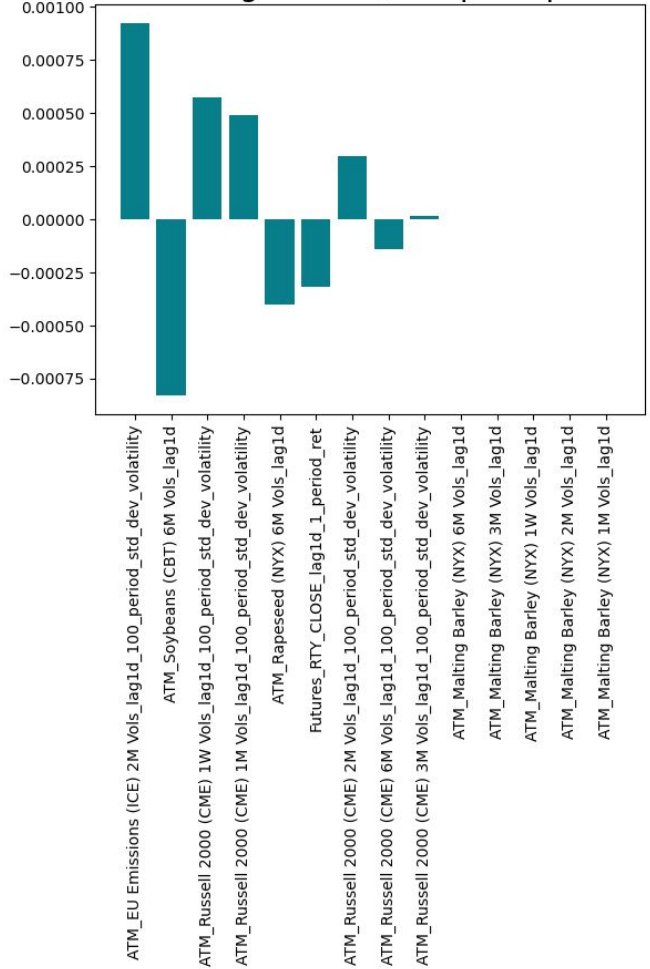
PCA: Explained Variance



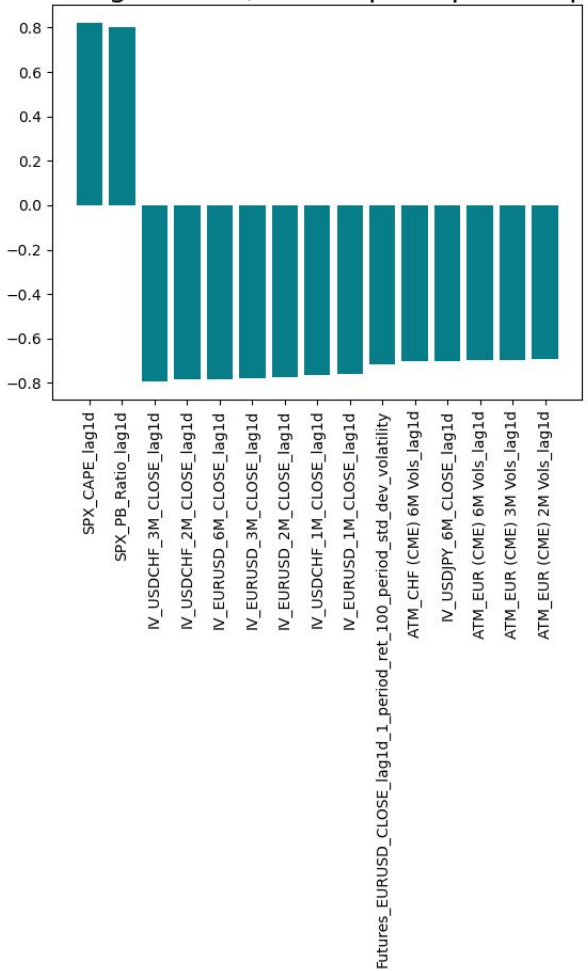
PCA loading scores (first principal component)



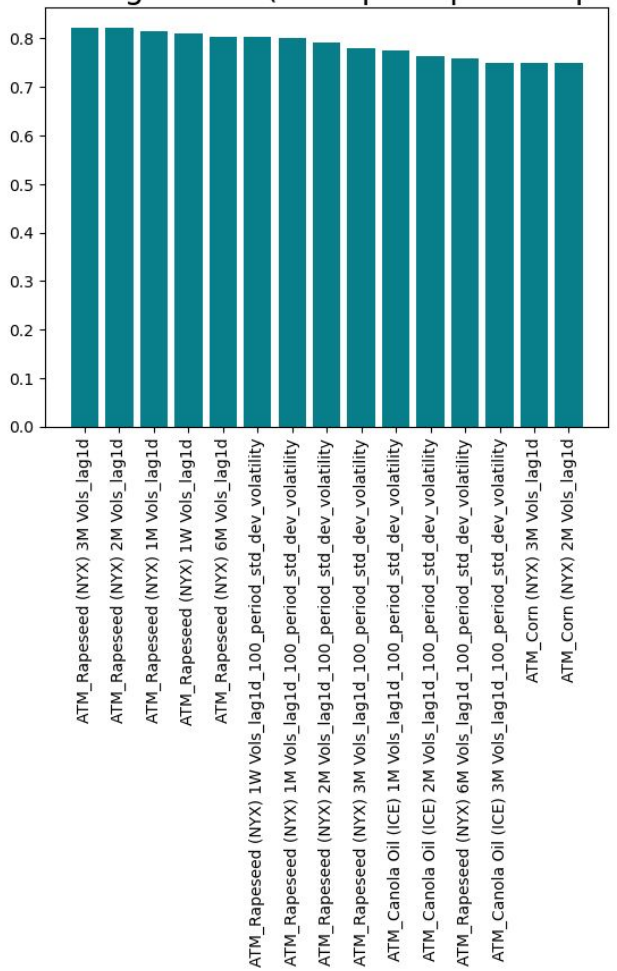
Bottom PCA loading scores (first principal component)



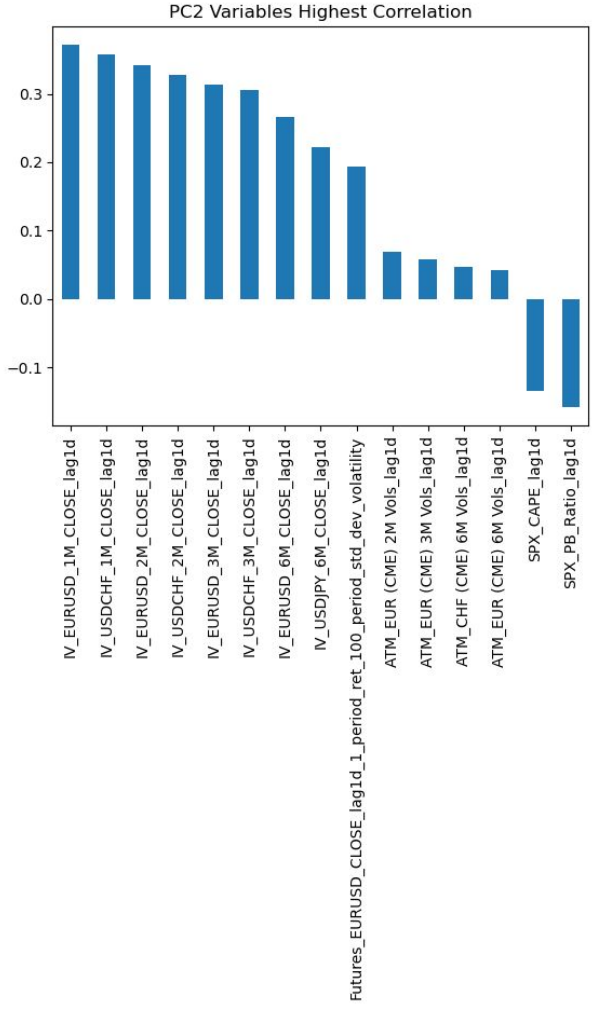
PCA loading scores (second principal component)



PCA loading scores (third principal component)

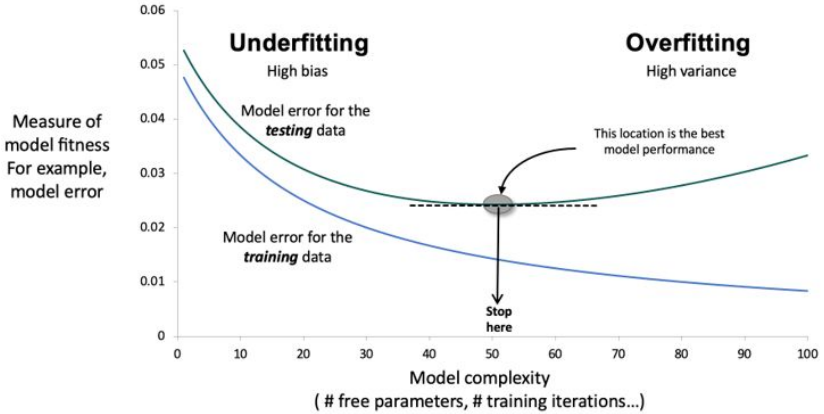



PC1 Variables Highest Correlation
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
ATM_S&P500 (CME) 3M Vols_lag1d_100_period_std_dev_volatility
ATM_S&P500 (CME) 2M Vols_lag1d_100_period_std_dev_volatility
IV_SPX_Call_ATM_30D_CLOSE_lag1d_100_period_std_dev_volatility
ATM_Nasdaq 100 (CME) 2M Vols_lag1d_100_period_std_dev_volatility
ATM_Nasdaq 100 (CME) 1M Vols_lag1d_100_period_std_dev_volatility
ATM_S&P500 (CME) 1M Vols_lag1d_100_period_std_dev_volatility
ATM_Nasdaq 100 (CME) 3M Vols_lag1d_100_period_std_dev_volatility
ATM_S&P500 (CME) 1W Vols_lag1d_100_period_std_dev_volatility
IV_RUT_Call_ATM_30D_CLOSE_lag1d_100_period_std_dev_volatility
ATM_30Y Bond (CBT) 1M Vols_lag1d_100_period_std_dev_volatility
ATM_Nasdaq 100 (CME) 1W Vols_lag1d_100_period_std_dev_volatility
IV_SPX_Call_ATM_60D_CLOSE_lag1d_100_period_std_dev_volatility
ATM_S&P500 (CME) 6M Vols_lag1d_100_period_std_dev_volatility
IV_RUT_Call_ATM_60D_CLOSE_lag1d_100_period_std_dev_volatility
IV_SPX_Call_ATM_90D_CLOSE_lag1d_100_period_std_dev_volatility


PC2 Variables Highest Correlation
0.3
0.2
0.1
0.0
−0.1
IV_EURUSD_1M_CLOSE_lag1d
IV_USDCHF_1M_CLOSE_lag1d
IV_EURUSD_2M_CLOSE_lag1d
IV_USDCHF_2M_CLOSE_lag1d
IV_EURUSD_3M_CLOSE_lag1d
IV_USDCHF_3M_CLOSE_lag1d
IV_EURUSD_6M_CLOSE_lag1d
IV_USDJPY_6M_CLOSE_lag1d
Futures_EURUSD_CLOSE_lag1d_1_period_ret_100_period_std_dev_volatility
ATM_EUR (CME) 2M Vols_lag1d
ATM_EUR (CME) 3M Vols_lag1d
ATM_CHF (CME) 6M Vols_lag1d
ATM_EUR (CME) 6M Vols_lag1d
SPX_CAPE_lag1d
SPX_PB_Ratio_lag1d

Methodology for Model Selection

- After completing the feature selection (PCA, Autoencoder …), we will run several models to find out the best model for prediction
- Starting from linear model as baseline and moving deeper to non-linear models and deep learning models.
- For each type of models, fine-tuning the model complexity to avoid underfitting and overfitting, as the graph shown
- For complicated models that have randomness, run them multiple times and using the average as model performance.



Model Selection: Linear Models

- Linear models are robust, hard to make mistakes. Nonlinear models are easy to make mistakes, especially when using complicated model

- A linear model result gives a baseline. Nonlinear models should always do at least as good as a linear model

- $Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp+\epsilon$, where $X1, X2, \ldots, Xp$ represent the multiple independent variables, $\beta 0, \beta 1, \beta 2, \ldots, \beta p$ are the parameters to be estimated

Model Selection: Non-Linear Models

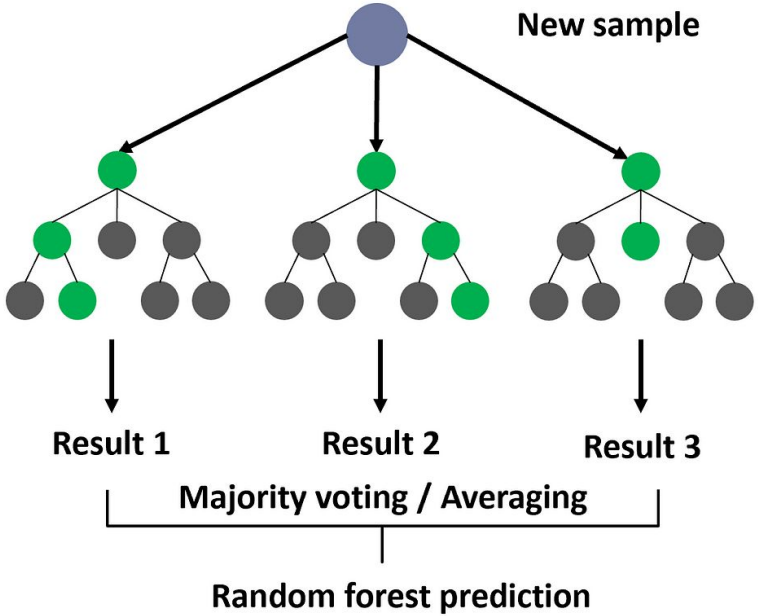
- There are many types of Nonlinear machine learning models. We want to focus on the models that fit the nature of our dataset (Financial Data, Time Series Components, High Randomness).

- Start with minimal complexity, increase complexity until overfitting (training substantially better than testing). Then back off to choose the lowest complexity that gives sufficiently good testing results.

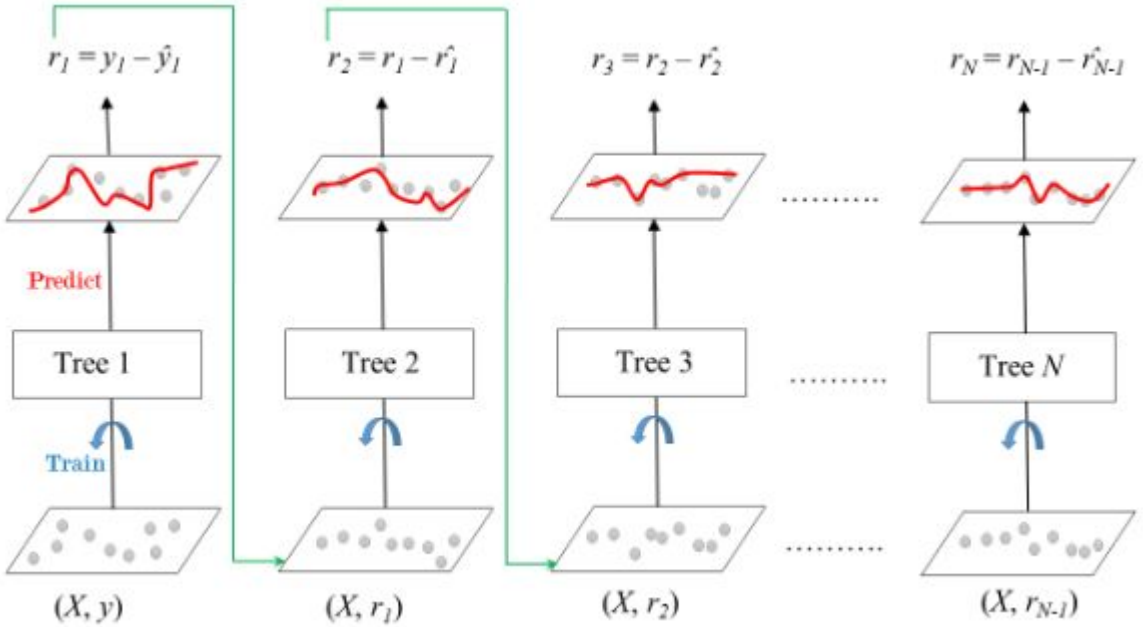
- Applying the method above on several models and listing the performances. Evaluating the model based on their accuracy, robustness, and interpretability… .

- **Potential Models:**
  - Random Forest
  - Gradient Boosting
  - Support Vector Regression

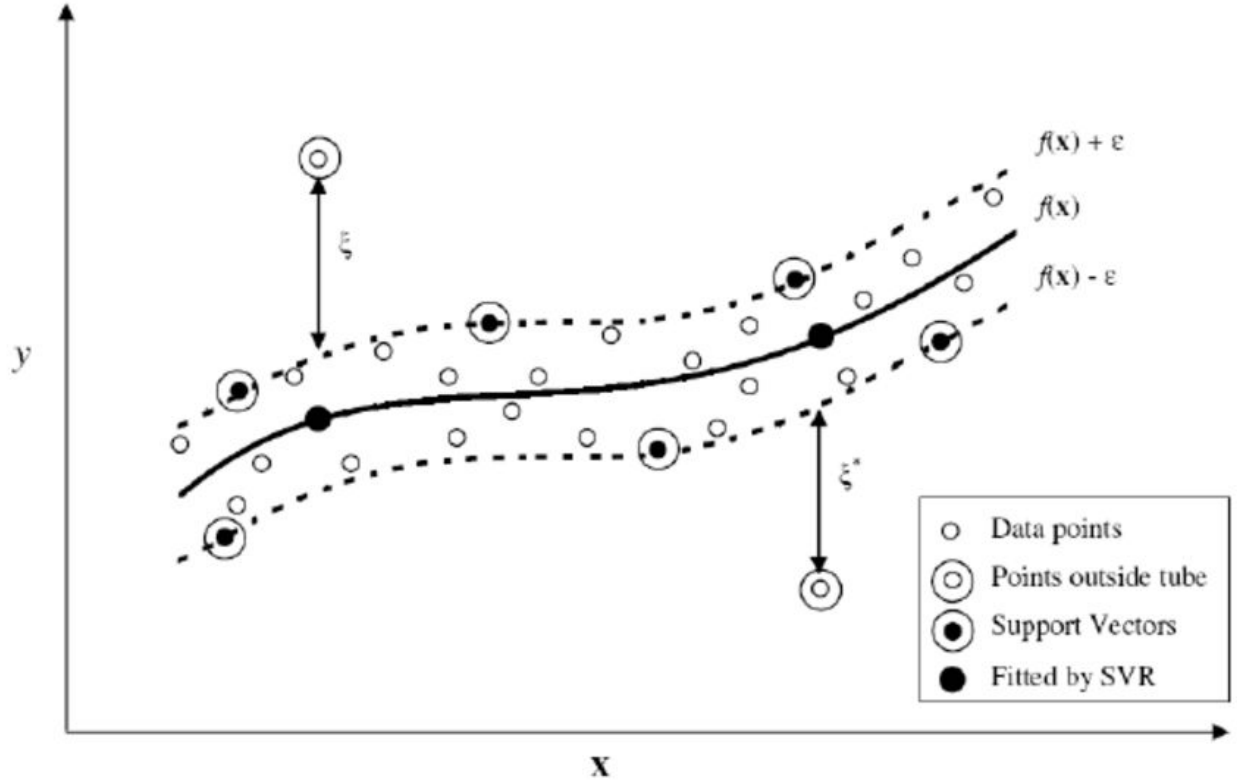
Random Forest:



Gradient Boosting:



SVR:


f(x) + ε
f(x)
f(x) - ε
ξ
ξ*
y
x
Data points
Points outside tube
Support Vectors
Fitted by SVR

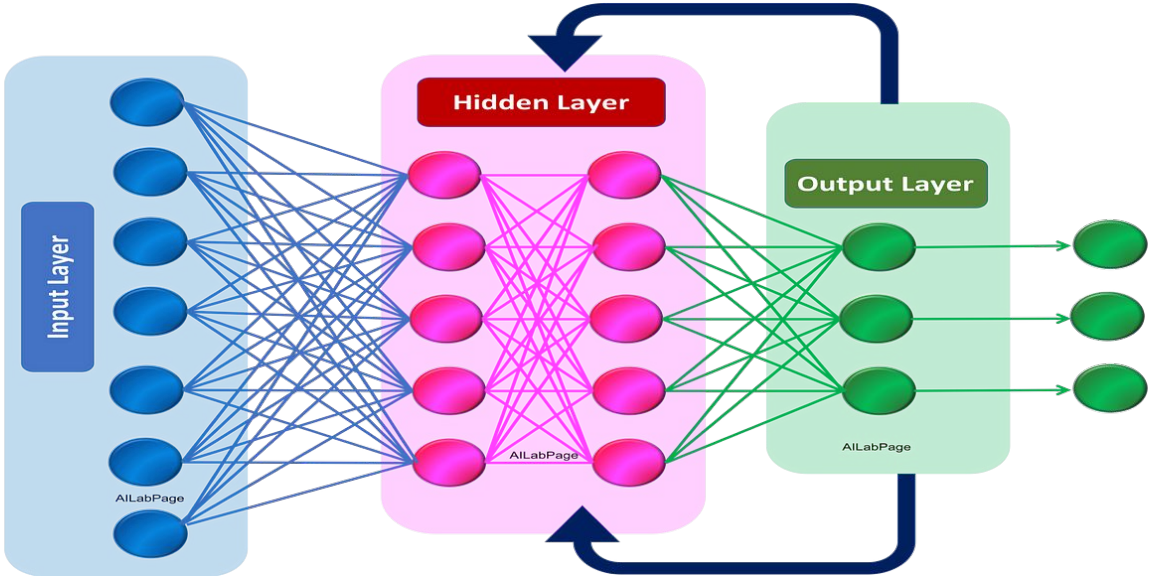
Model Selection: Deep Learning

- Deep Learning models are more complex non-linear models. We will focus on Neural Network models. The approach of creating and evaluating deep learning models is similar to non-linear models.

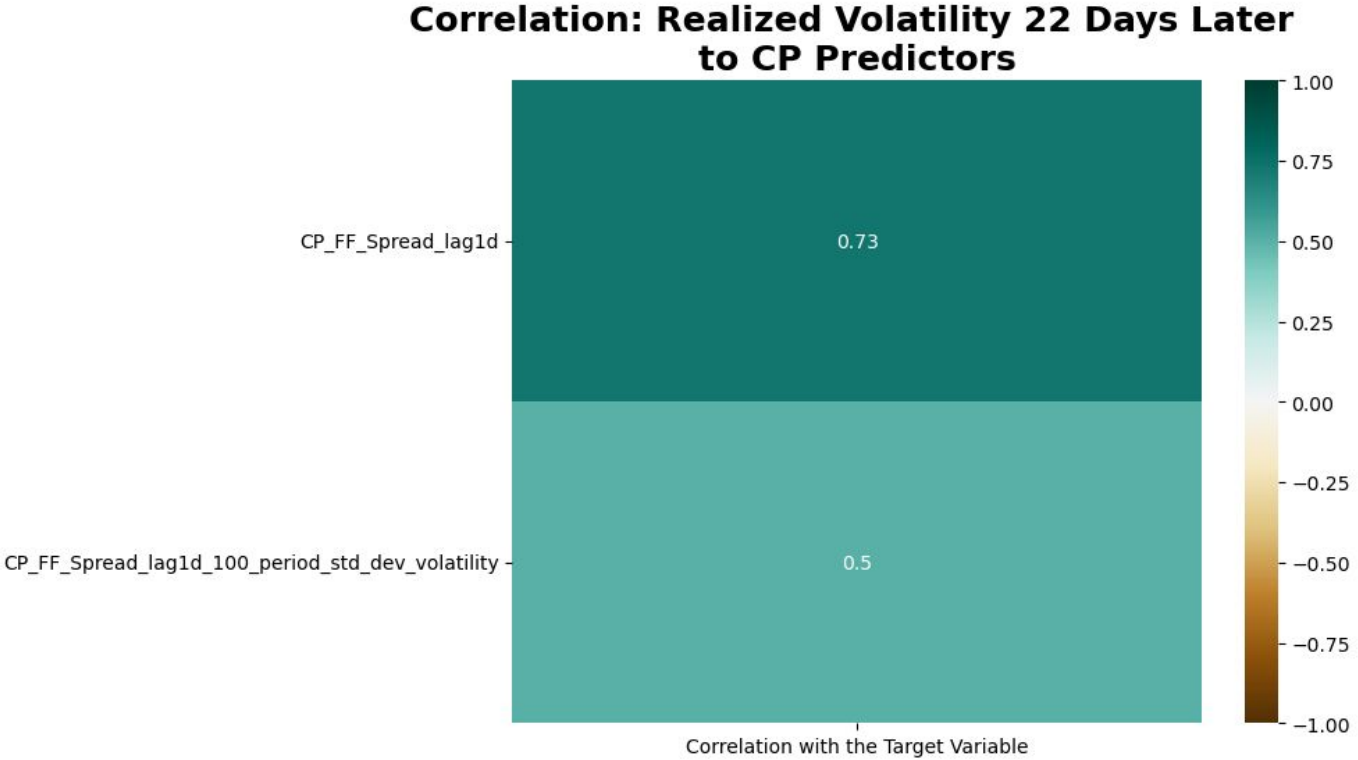
- Time efficiency might be a big limitation when modeling by deep learning model. Considering use GPU CUDA will increase the speed of training, especially for larger and more complex deep learning architectures.

- **Potential Model:**
  - Neural Networks(NN) & Recurrent Neural Networks (RNN)
  - Long Short-Term Memory & Gated Recurrent Unit (Variants of RNN)

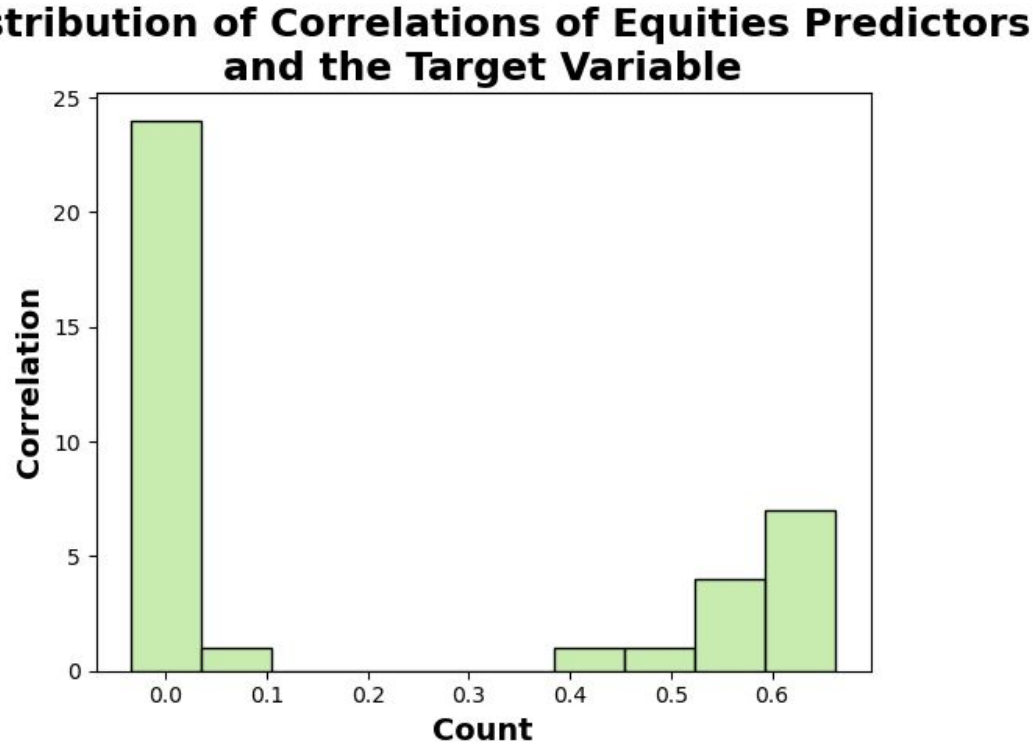
Recurrent Neural Networks:



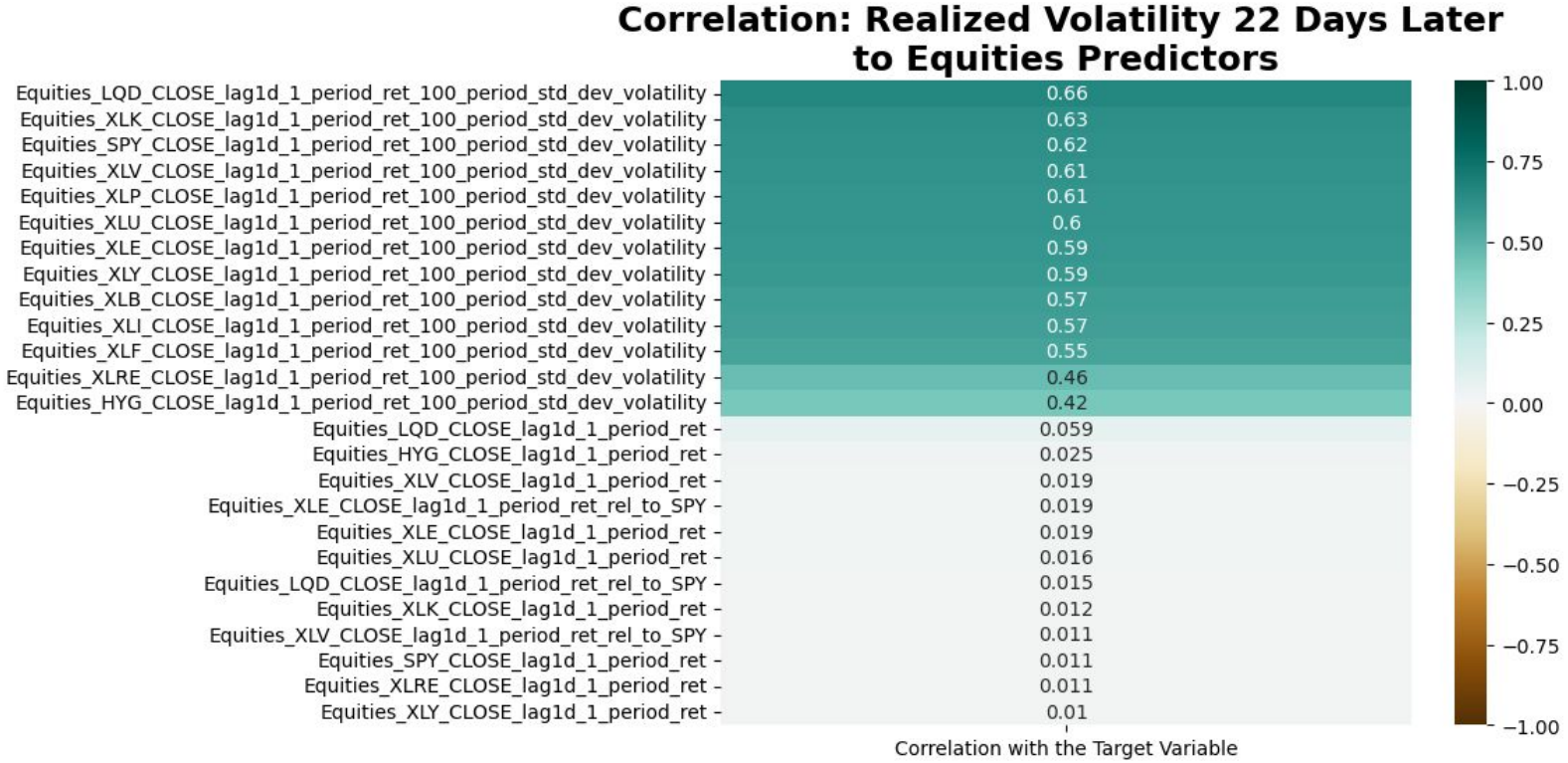
CP Predictors: Highly Correlated



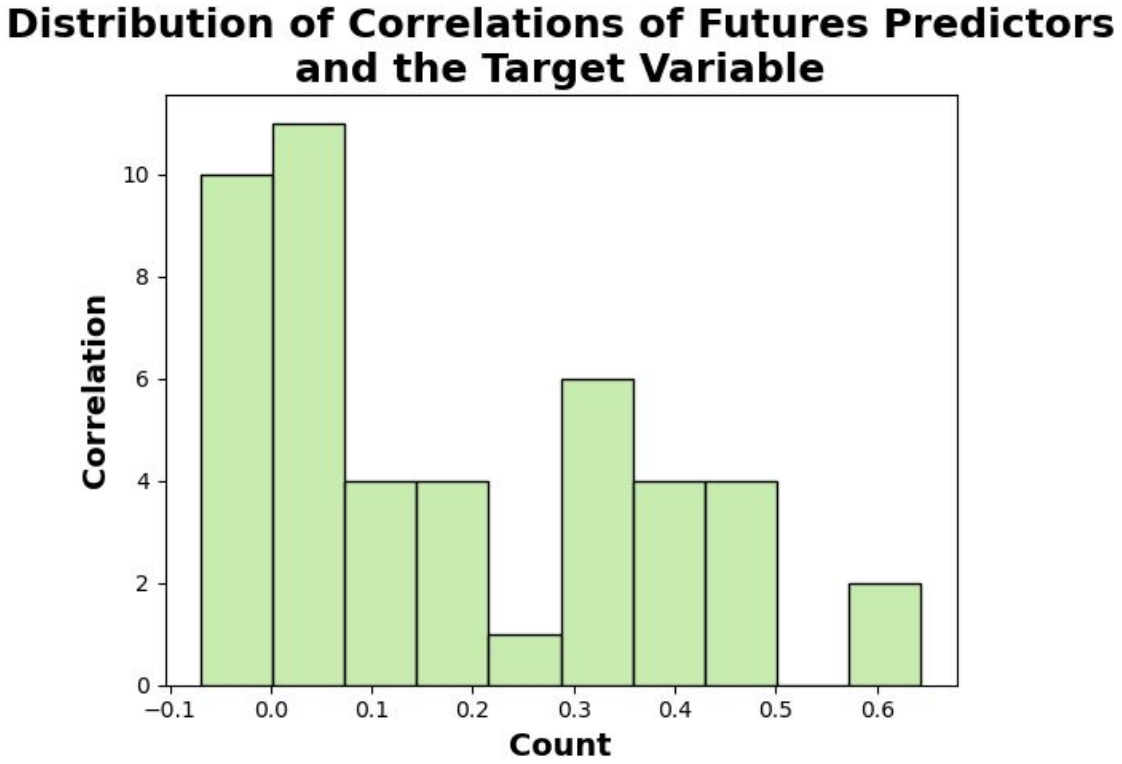
Equities Predictors:



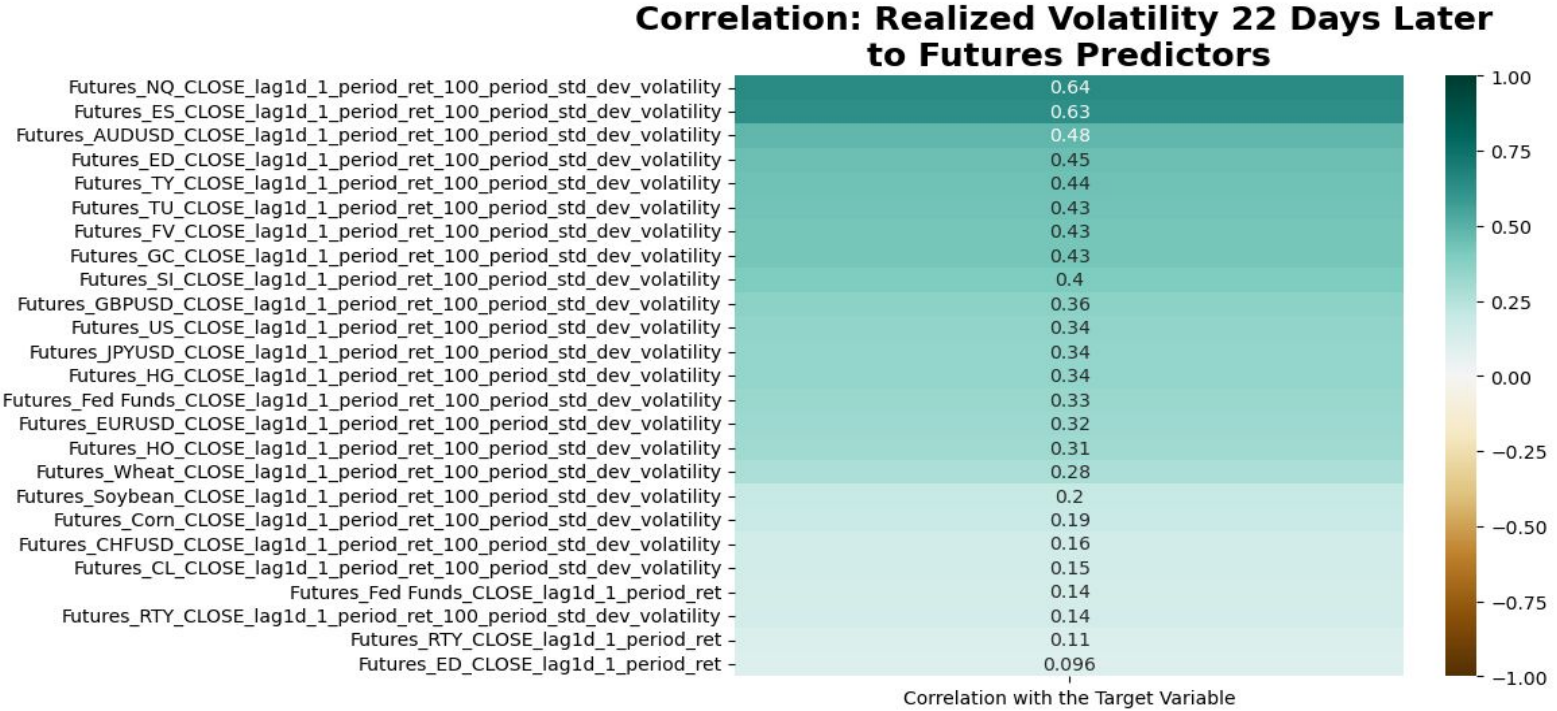
Equities Predictors:



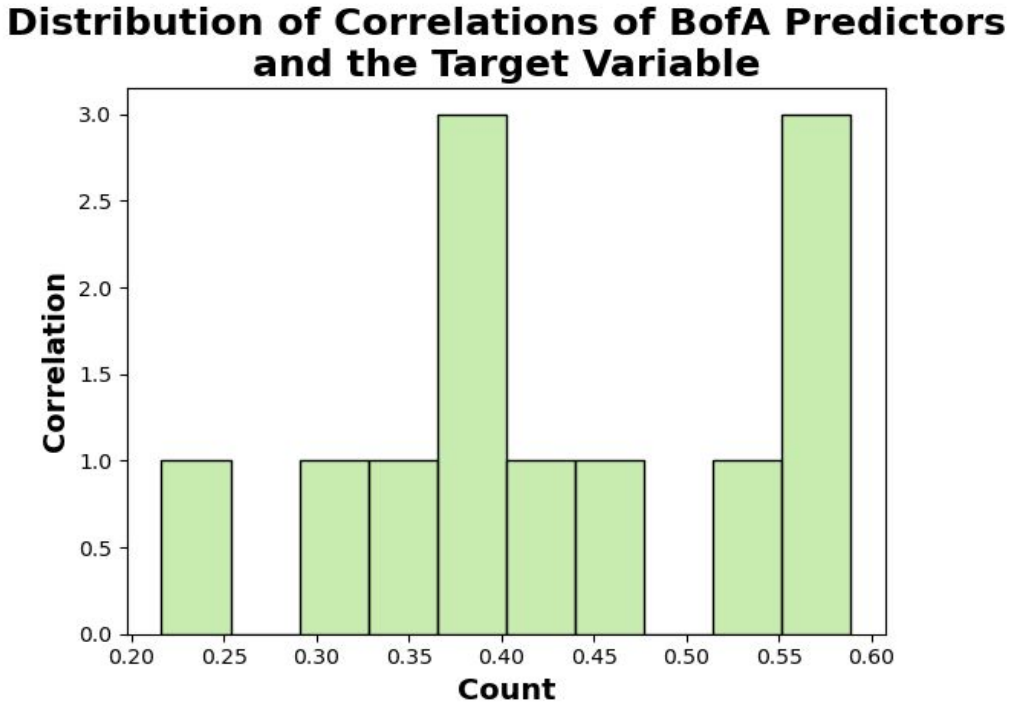
Futures Predictors:



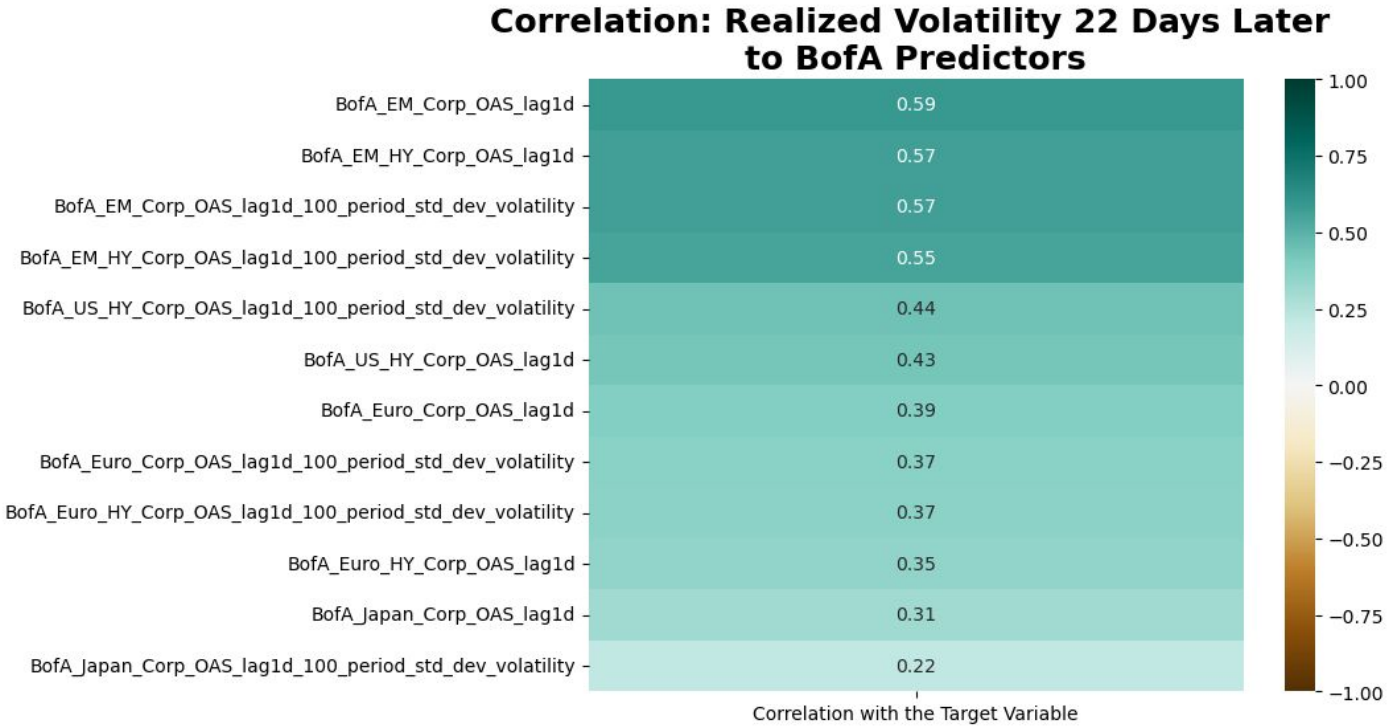
Futures Predictors:



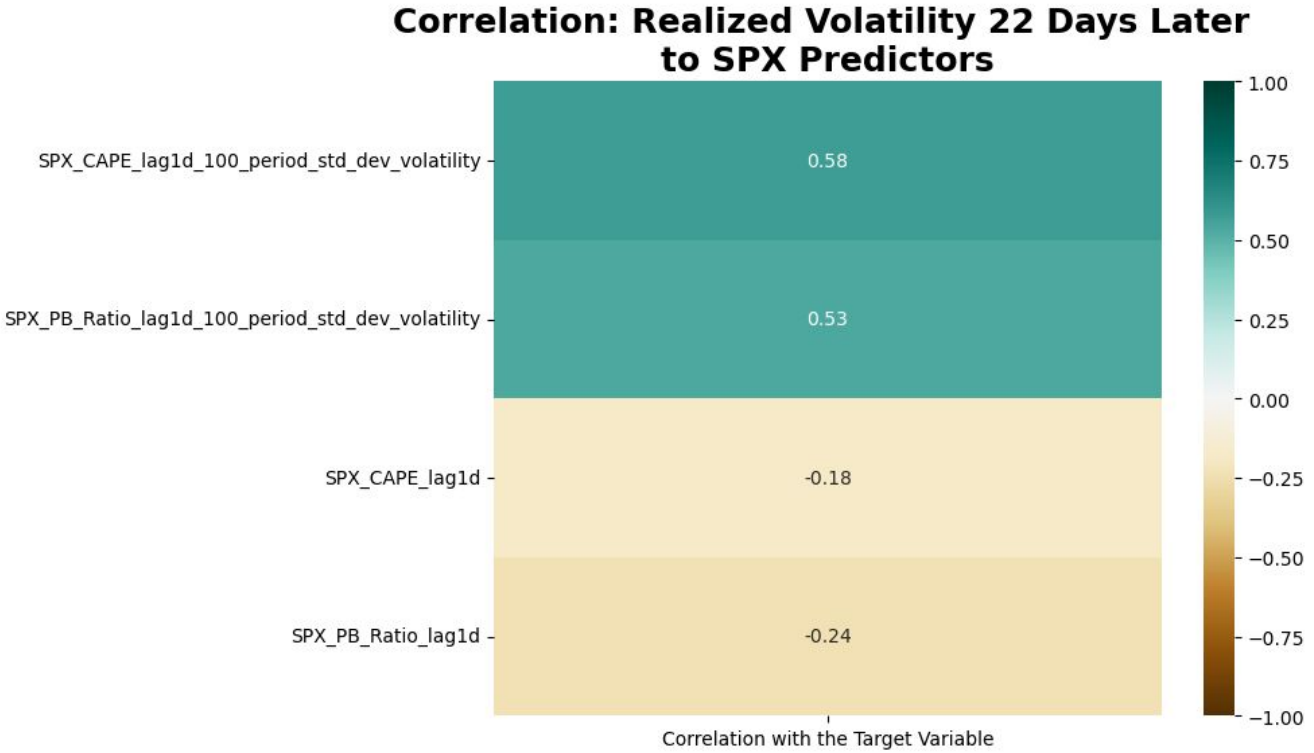
BofA Predictors: Fairly Correlated



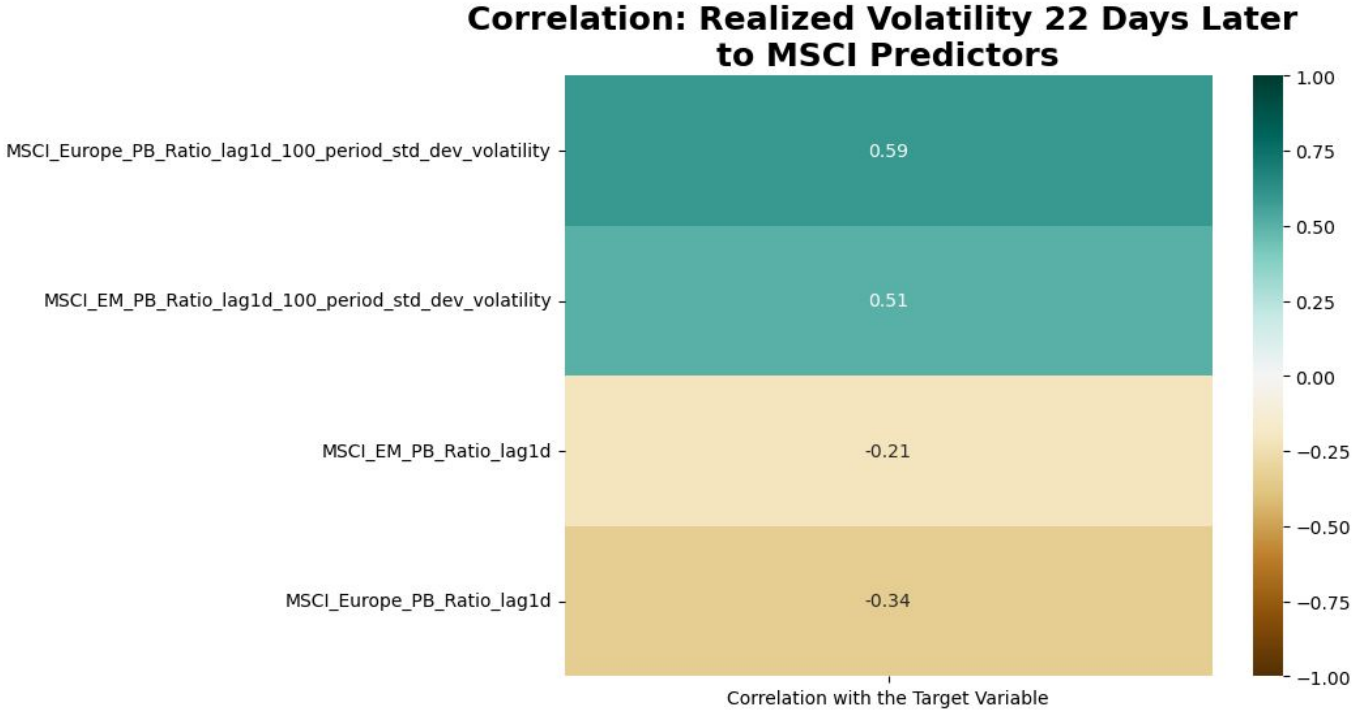
BofA Predictors: Fairly Correlated:



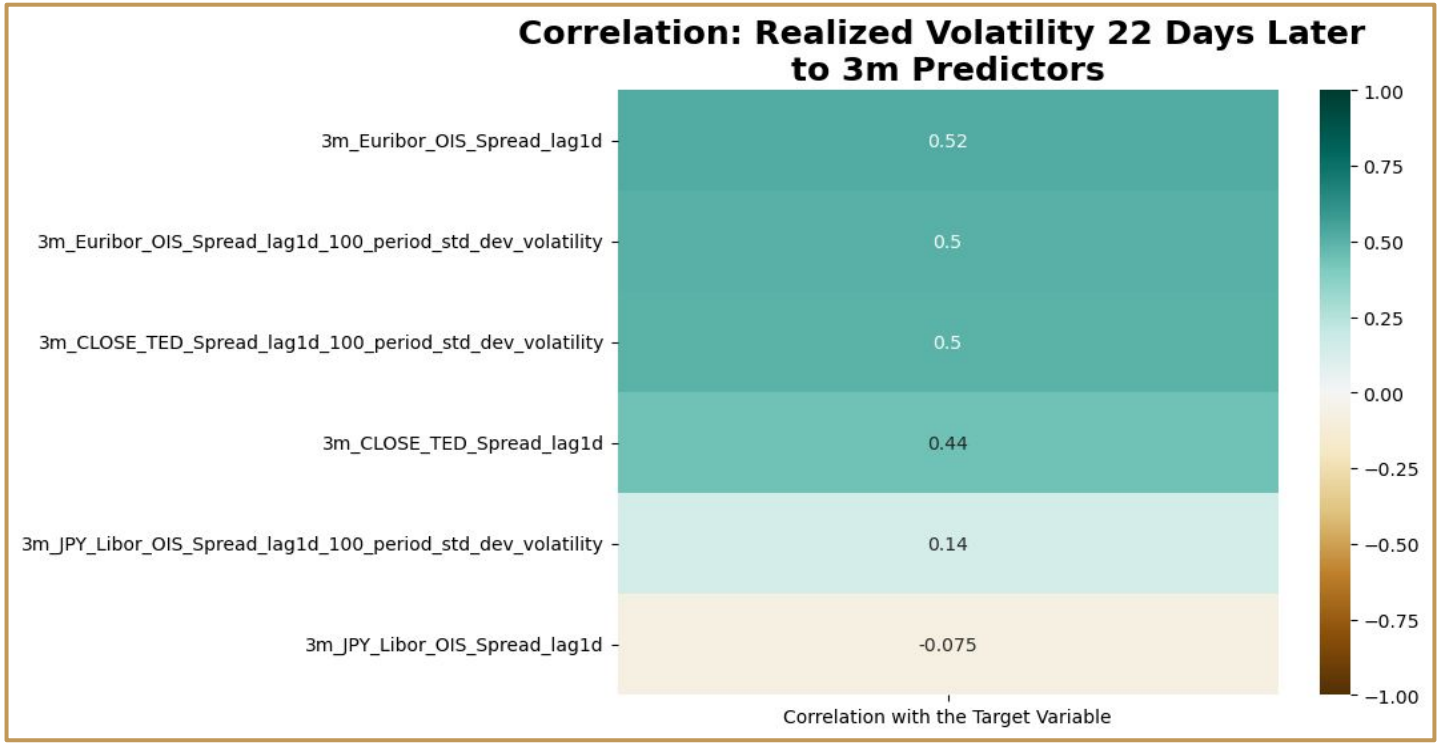
SPX Predictors:



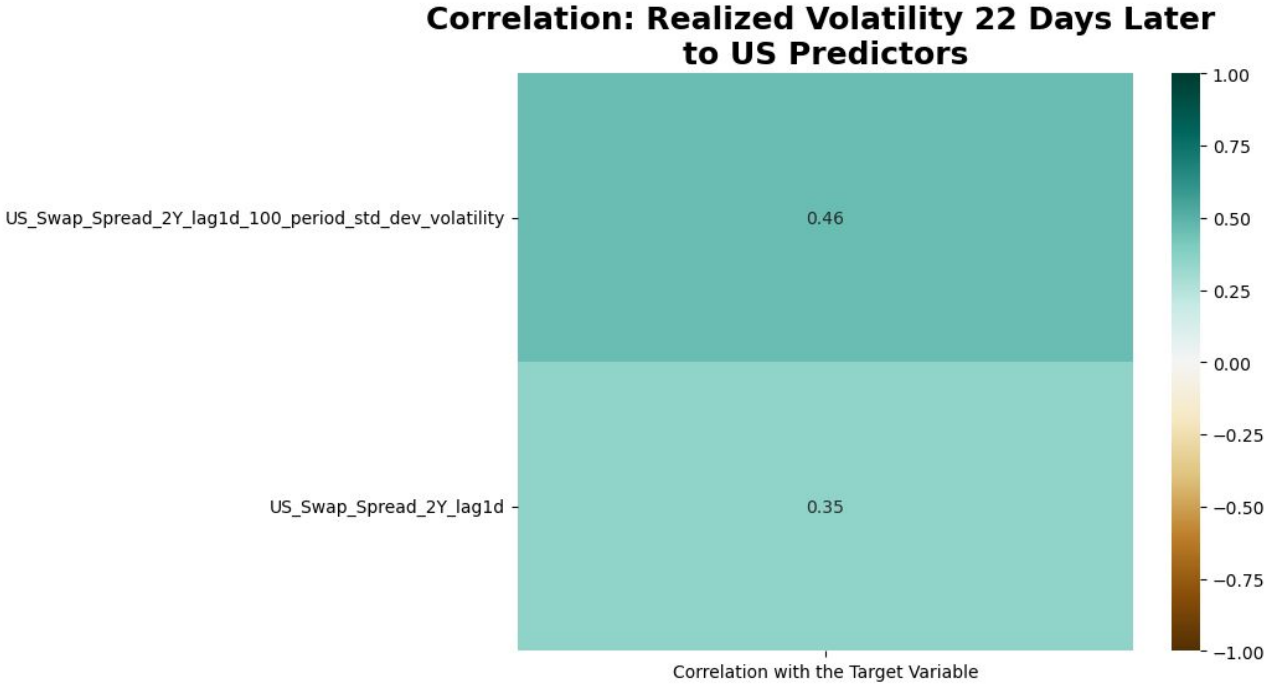
MSCI Predictors:



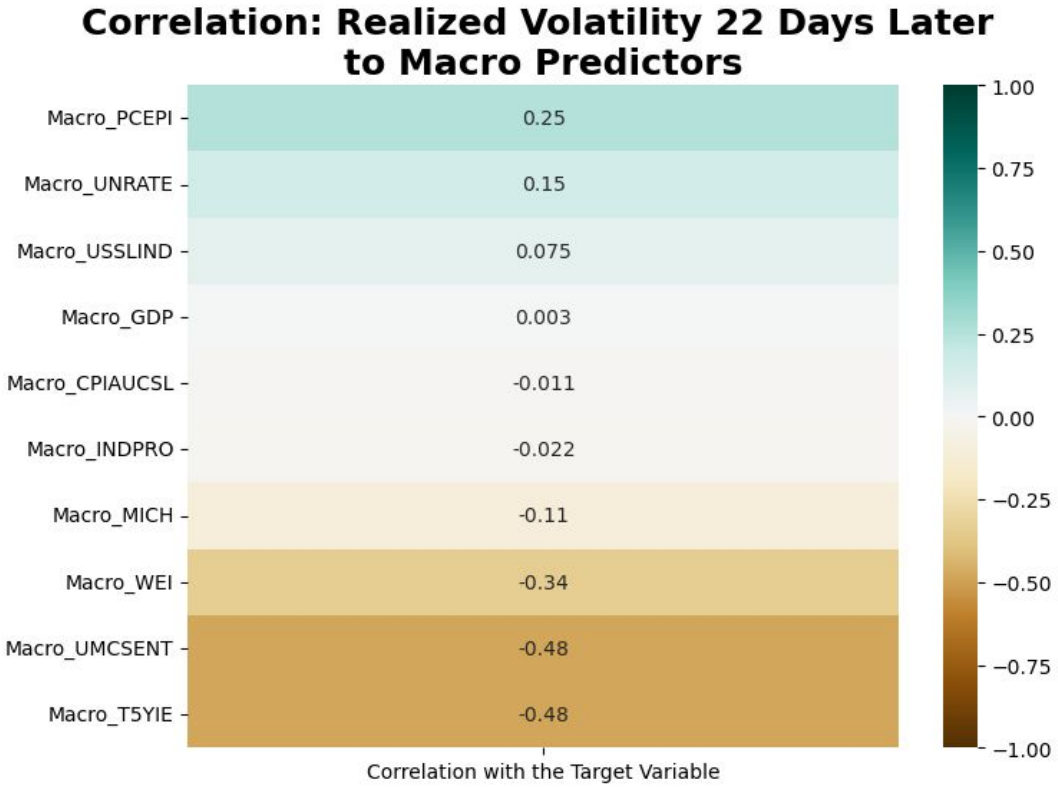
3M Predictors



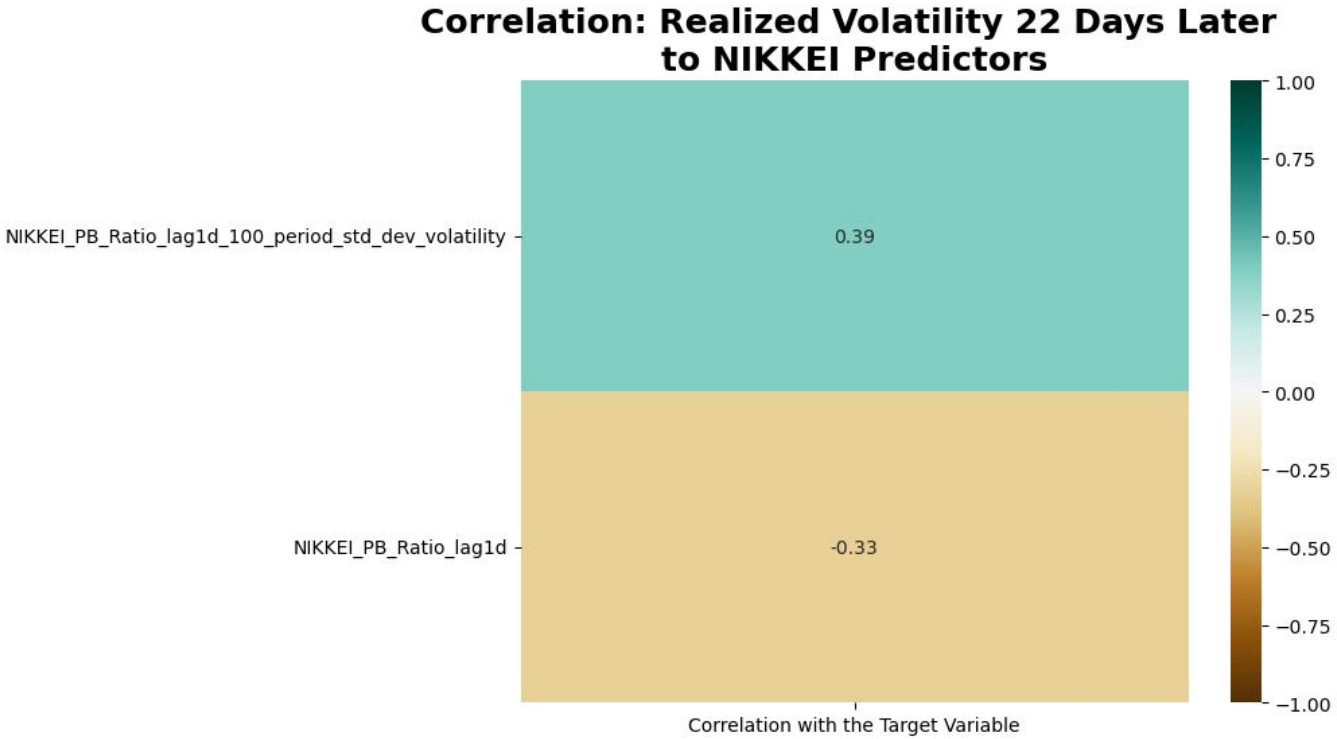
US Swap Spread Predictors:



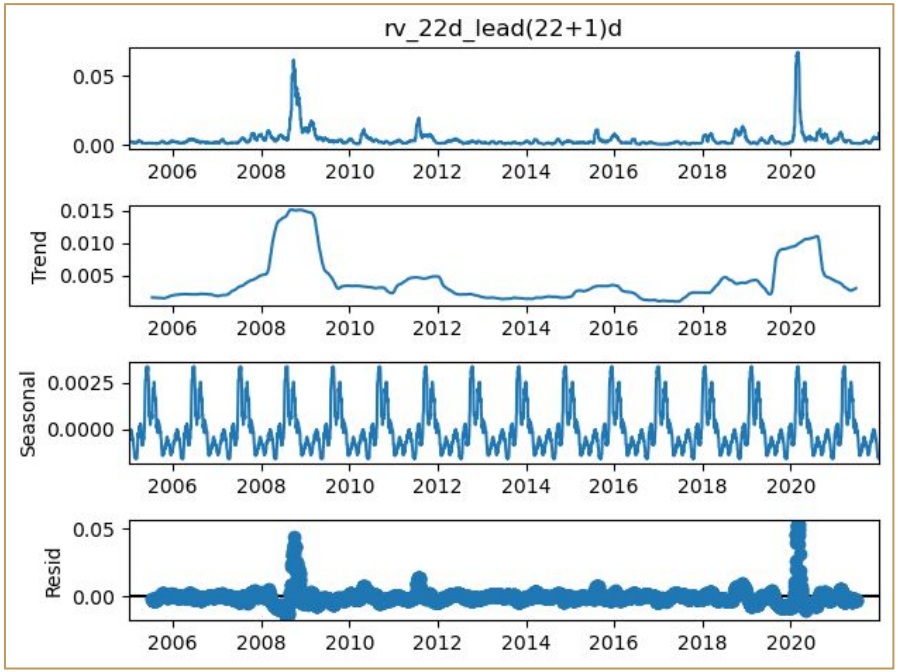
Macro Predictors:

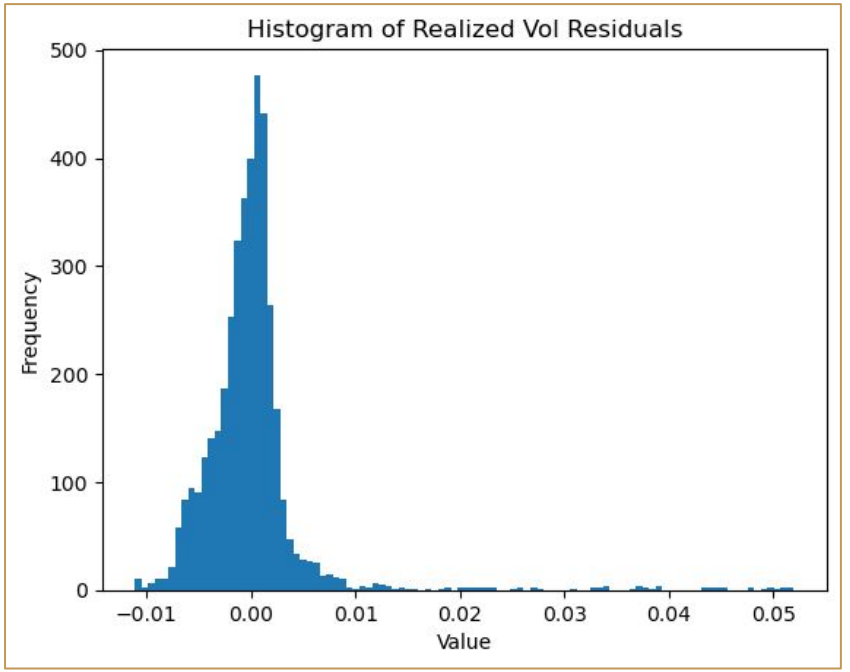


NIKKEI Predictors:



About the Data





Thank you!