Developing a Macro Factor Investing Model

Internship presentation - Sanskriti Bajaj

Steps followed to develop the Macro Factor Investing Model

1. Understanding Factor Investing
2. Data Research and Aggregation
3. Data Analysis
4. Regression Analysis
5. Cross Validation Analysis

1. Understanding Factor Investing

What is Factor Investing?

Factor investing is an investment approach that involves targeting specific drivers of return across asset classes

There are two main types of factors:

- **Macroeconomic factors** - capture broad risks across asset classes
- **Style factors** - help in explaining returns and risk within asset classes

Macroeconomic Factors

- **Economic growth** -Surprise in GDP growth
- **Real rates** - Surprise in real interest rates
- **Inflation** - Surprise in inflation levels
- **Emerging Markets** - Spread between emerging and developed market securities
- **Credit** - Surprise in default rates
- **Liquidity** - Spread between small-cap and large-cap equities

Style Factors

- **Value** - stocks that are cheaper relative to their fundamental value
- **Minimum Volatility** - Stocks that historically exhibit low volatility
- **Momentum** - Stocks with an upward price trend
- **Quality** - Companies with healthy financial statements
- **Size** - Companies with smaller market capitalization
- **Dividend yield** - Companies with higher dividend yield relative to price

2. Data Research and Aggregation

Data Research

- Analyzed the following macroeconomic factors -
  - Nominal GDP and Real GDP
  - 10 Year G-Sec yields and 5 Year G-Sec yields
  - CPI and WPI
  - Foreign Investment Inflows
  - Fiscal Deficit
  - M1, M2 and M3
  - Nifty Commodity Index
  - Nominal Exchange Rate (NEER) and Real Exchange Rate (REER)
- Used the following data sources -
  - Thomson Reuters
  - RBI Economic Database
  - Ministry of Statistics and Programme Implementation (MOSPI)
  - Bank of International Settlements (BIS)

Data Aggregation

Aggregated the Data in a Multi-Index Pandas Dataframe using Jupyter Notebook

```
Data columns (total 11 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Quarter           4133 non-null   object
 1   Nifty_500_Return  4133 non-null   float64
 2   Commodity_Index   4133 non-null   float64
 3   NEER              4133 non-null   float64
 4   10_Year_Yield     4133 non-null   float64
 5   CPI               4133 non-null   float64
 6   Nominal_GDP       4133 non-null   float64
 7   M2                4133 non-null   float64
 8   FII_Inflows       4133 non-null   float64
 9   Fiscal_Deficit    4133 non-null   float64
 10  const             4133 non-null   int64
dtypes: float64(9), int64(1), object(1)
memory usage: 399.8+ KB
```

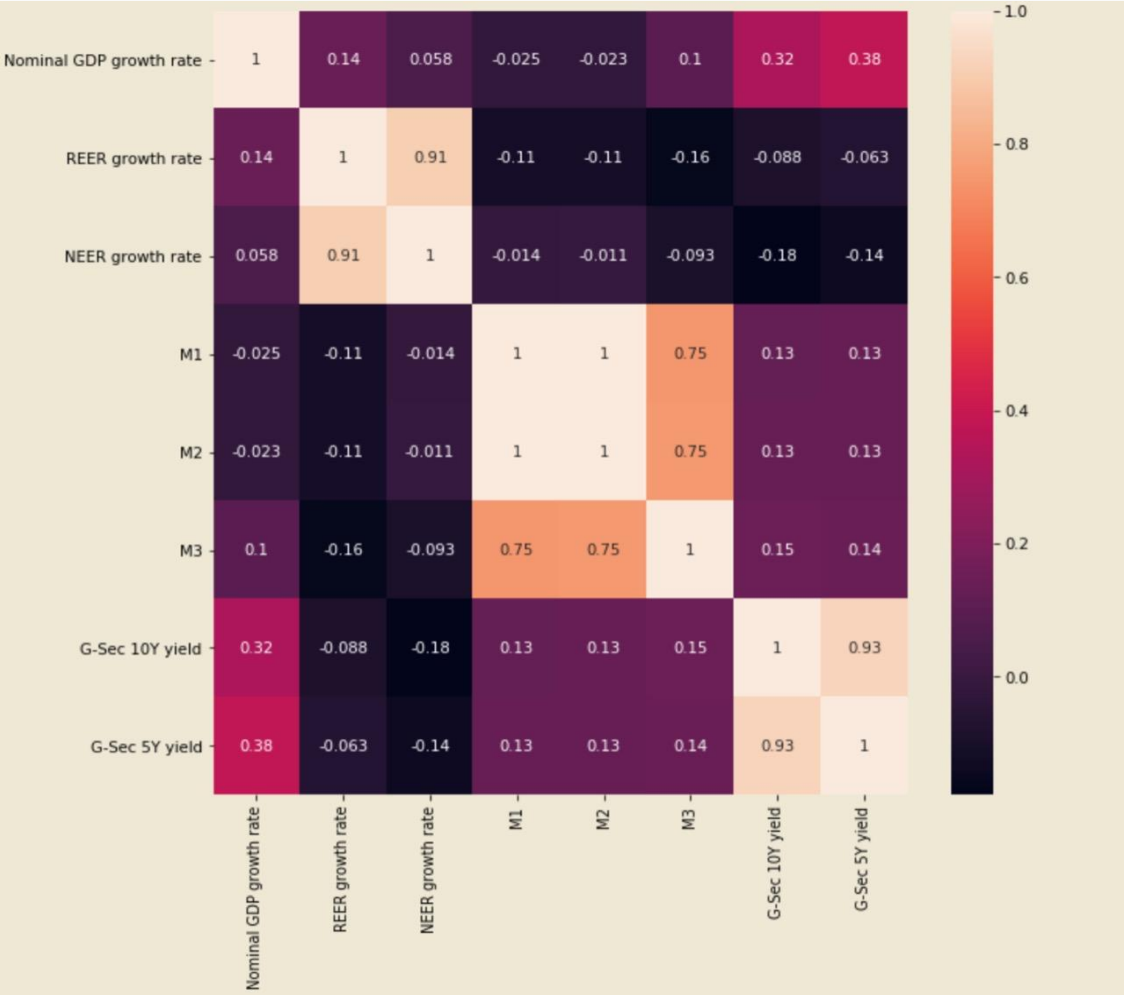
Year	Date	Quarter	Nifty_500_Return	Commodity_Index	NEER	10_Year_Yield	CPI	Nominal_GDP	M2	FII_Inflows
2020-21	2021-03-31	Q1-2021	-0.60	-0.07	0.30	0.55	-9.36	18.5	3.88	-114.96
	2021-03-30	Q1-2021	1.97	1.76	-0.94	0.33	-9.36	18.5	3.88	-114.96
	2021-03-26	Q1-2021	1.30	1.43	0.16	-0.13	-9.36	18.5	3.88	-114.96
	2021-03-25	Q1-2021	-1.67	-1.82	0.16	-0.34	-9.36	18.5	3.88	-114.96
	2021-03-24	Q1-2021	-1.77	-2.00	0.03	0.13	-9.36	18.5	3.88	-114.96
	2021-03-23	Q1-2021	0.64	0.61	0.04	-0.58	-9.36	18.5	3.88	-114.96
	2021-03-22	Q1-2021	0.20	1.02	0.21	-0.21	-9.36	18.5	3.88	-114.96
	2021-03-19	Q1-2021	1.15	2.39	0.32	-0.15	-9.36	18.5	3.88	-114.96
	2021-03-18	Q1-2021	-1.18	-0.79	-0.23	0.27	-9.36	18.5	3.88	-114.96
	2021-03-17	Q1-2021	-1.54	-2.41	0.03	0.08	-9.36	18.5	3.88	-114.96

3. Data Analysis

Data Analysis

Analyzed the correlations between different indicators to reduce the incidence of multicollinearity among explanatory variables

Utilized the seaborn library to generate a correlation matrix

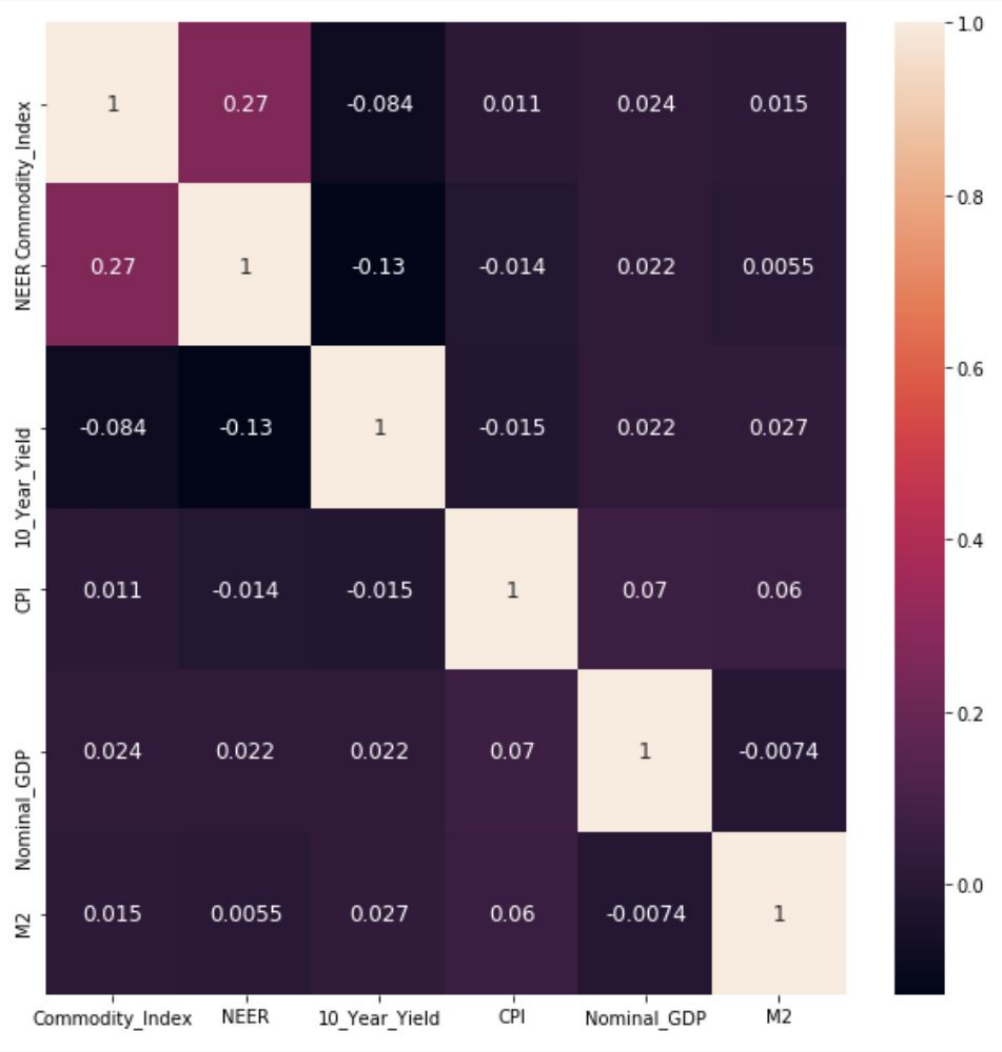


Data Analysis

The correlation between indicators of similar nature (for example M1, M2 and M3) was substantially high hence these were removed to avoid multicollinearity in the model

The following indicators were finally chosen to be included in the model as explanatory features -

1. NIFTY Commodity Index returns
2. Nominal exchange rate (NEER)
3. 10 Year G-Sec yields
4. Nominal GDP
5. M2
6. CPI



4. Regression Analysis

Regression

Ran PooledOLS regression for the panel data. Used Statmodel's PooledOLS module to generate regression results

The model parameters were successful in explaining 88% of the change in NIFTY 500 Index returns

PooledOLS Estimation Summary

Dep. Variable:	Nifty_500_Return	R-squared:	0.8813
Estimator:	PooledOLS	R-squared (Between):	0.8747
No. Observations:	4133	R-squared (Within):	0.8814
Date:	Tue, Jul 06 2021	R-squared (Overall):	0.8813
Time:	18:22:07	Log-likelihood	-2874.9
Cov. Estimator:	Unadjusted		
		F-statistic:	3827.6
Entities:	17	P-value	0.0000
Avg Obs:	243.12	Distribution:	F(8,4124)
Min Obs:	237.00		
Max Obs:	249.00	F-statistic (robust):	3827.6
		P-value	0.0000
Time periods:	4133	Distribution:	F(8,4124)
Avg Obs:	1.0000		
Min Obs:	1.0000		
Max Obs:	1.0000		

Parameter Estimates

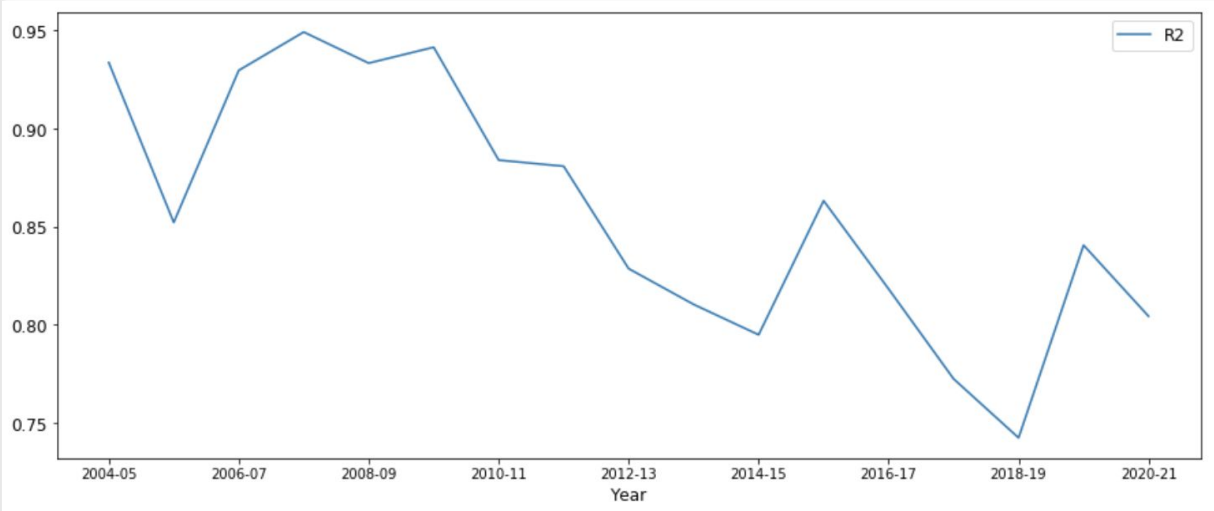
	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0233	0.0170	1.3706	0.1706	-0.0100	0.0565
Commodity_Index	0.7971	0.0048	165.36	0.0000	0.7876	0.8065
NEER	0.1568	0.0197	7.9736	0.0000	0.1183	0.1954
10_Year_Yield	-0.0450	0.0096	-4.6825	0.0000	-0.0638	-0.0261
CPI	0.0001	0.0005	0.2787	0.7805	-0.0008	0.0011
Nominal_GDP	-0.0002	0.0012	-0.1964	0.8443	-0.0026	0.0021
M2	0.0010	0.0039	0.2417	0.8090	-0.0068	0.0087
FII_Inflows	-3.538e-07	1.614e-06	-0.2192	0.8265	-3.518e-06	2.811e-06
Fiscal_Deficit	-2.161e-05	1.498e-05	-1.4427	0.1492	-5.098e-05	7.756e-06

Analysis of R2

Generated a visualization of R2 over the years to examine model robustness across time

The model demonstrated significantly high R2 values over time. R2 remained between 0.95 to 0.75

Model robustness was lowest (0.75) during the initial phase of the pandemic owing to significant deviation in Index returns as compared to past data

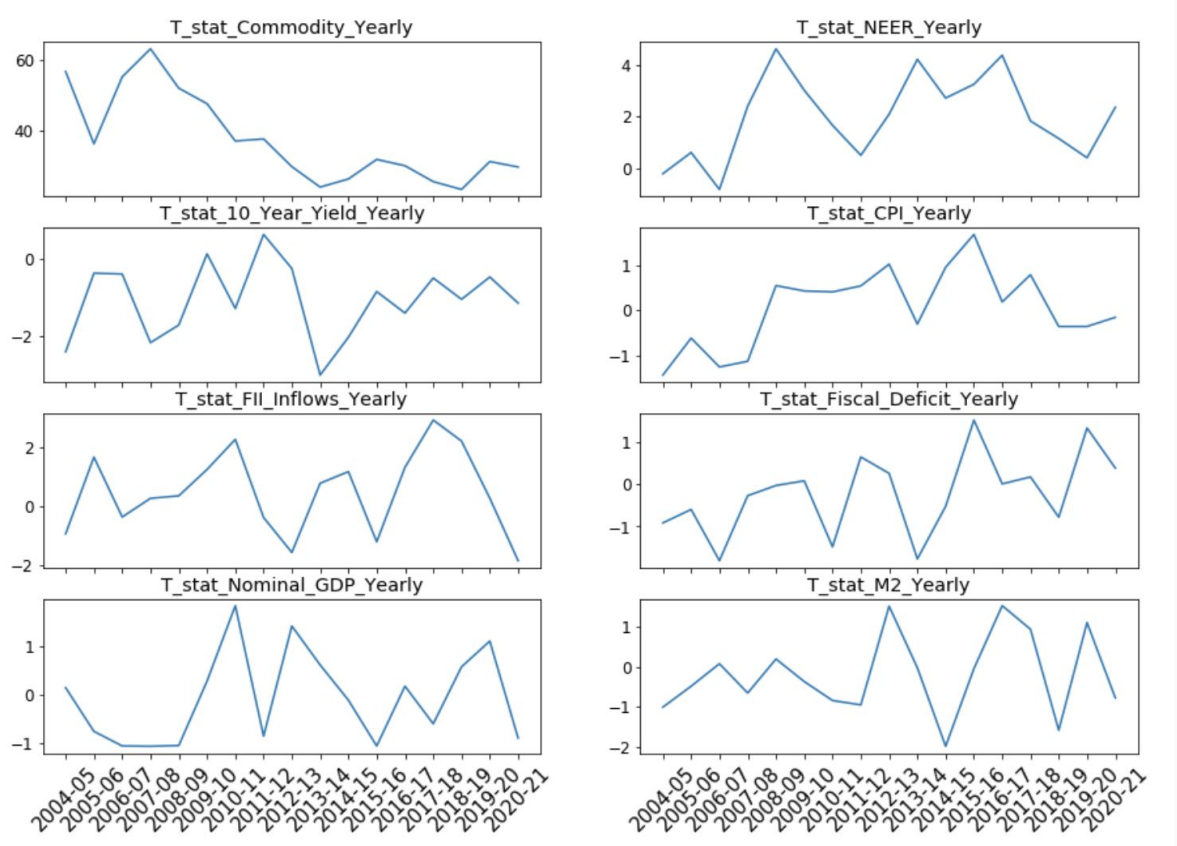


Analysis of T-stat values

Generated a visualization of T-stat values over time for different model features

The goal of this analysis was to identify parameters that were most successful in explaining changes in the target variable i.e. change in NIFTY 500 Index returns

The visualization demonstrated that the most significant explanatory variables were Commodity Index returns and Nominal exchange rate



5. Cross Validation Analysis

What is Cross Validation?

After a model has been fit to the entire dataset, the next step is to validate the model. Validation checks the model robustness beyond the usual measures of R2 and T-stat. It ensures that the model is able to make sound predictions across all time periods.

Process of validation involves dividing the data into different parts, some parts of the data are reserved for training the model and the other parts are used for testing the model predictions. The process flow is mentioned below -

1. Divide the data into test and train sets
2. Use the train dataset to fit the model
3. Use the test dataset to generate the model predictions
4. Compare the model predictions to the actual data

As part of this project I used two techniques for performing cross validation, 50-50 test train split validation and Time Series split validation.

50-50 test train split

For performing 50-50 split validation, data was split in half, the first half was used as train data

The idea behind 50-50 split validation is to use past data to predict future behaviour of target variables

PooledOLS regression was performed on the train data, the R2 value of 0.9216 highlighted the robustness of the regression model

PooledOLS Estimation Summary

Dep. Variable:	Nifty_500_Return	R-squared:	0.9216
Estimator:	PooledOLS	R-squared (Between):	0.8702
No. Observations:	2067	R-squared (Within):	0.9219
Date:	Wed, Jul 07 2021	R-squared (Overall):	0.9216
Time:	13:44:33	Log-likelihood	-1358.5
Cov. Estimator:	Unadjusted		
		F-statistic:	3022.5
Entities:	9	P-value	0.0000
Avg Obs:	229.67	Distribution:	F(8,2058)
Min Obs:	117.00		
Max Obs:	249.00	F-statistic (robust):	3022.5
		P-value	0.0000
Time periods:	2067	Distribution:	F(8,2058)
Avg Obs:	1.0000		
Min Obs:	1.0000		
Max Obs:	1.0000		

Parameter Estimates

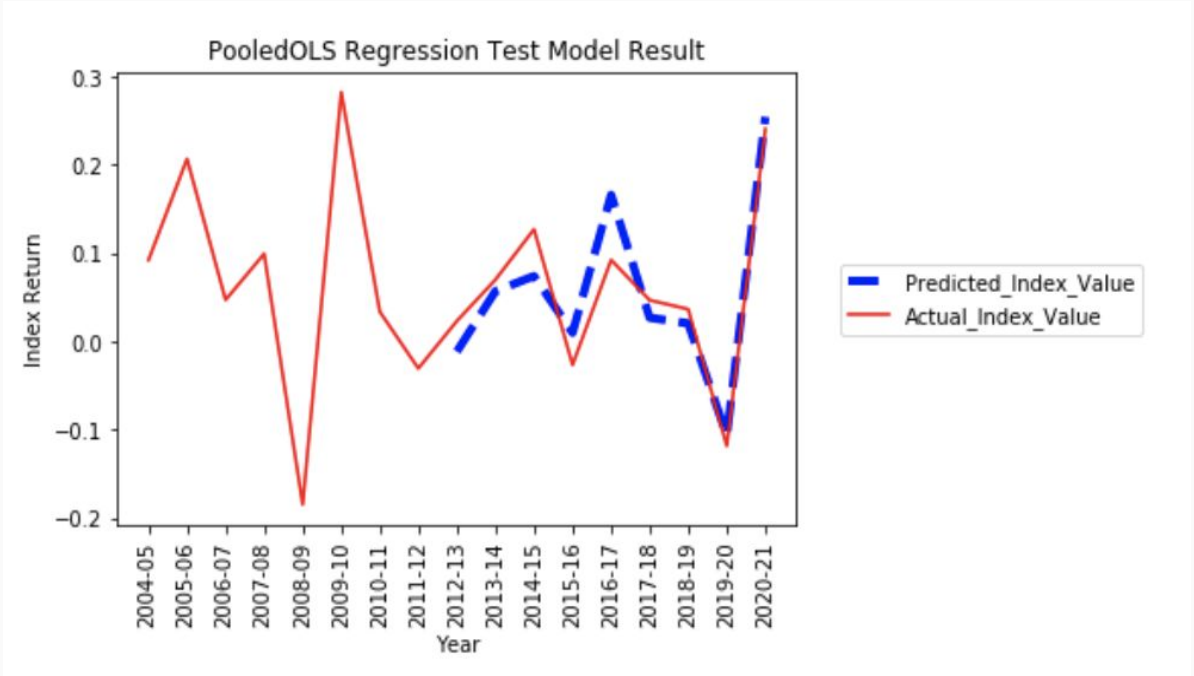
	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	0.0385	0.0398	0.9673	0.3335	-0.0395	0.1164
Commodity_Index	0.8332	0.0056	148.26	0.0000	0.8222	0.8442
NEER	0.1586	0.0262	6.0552	0.0000	0.1072	0.2099
10_Year_Yield	-0.0462	0.0119	-3.8746	0.0001	-0.0696	-0.0228
CPI	-0.0008	0.0008	-0.9465	0.3440	-0.0024	0.0008
Nominal_GDP	-0.0008	0.0025	-0.3184	0.7502	-0.0057	0.0041
M2	-0.0131	0.0073	-1.7919	0.0733	-0.0274	0.0012
FII_Inflows	-8.926e-07	1.609e-06	-0.5548	0.5791	-4.048e-06	2.263e-06
Fiscal_Deficit	-3.691e-05	2.413e-05	-1.5296	0.1263	-8.423e-05	1.041e-05

50-50 test train split

After fitting the model on the training data, the testing data was used for generating predictions

The predicted values were plotted along with the actual index values to visualize the performance of the model

The plot demonstrates that the predicted index values mimicked the actual index values for most years while there was deviation during years such as 2016-17 and 2013-14.

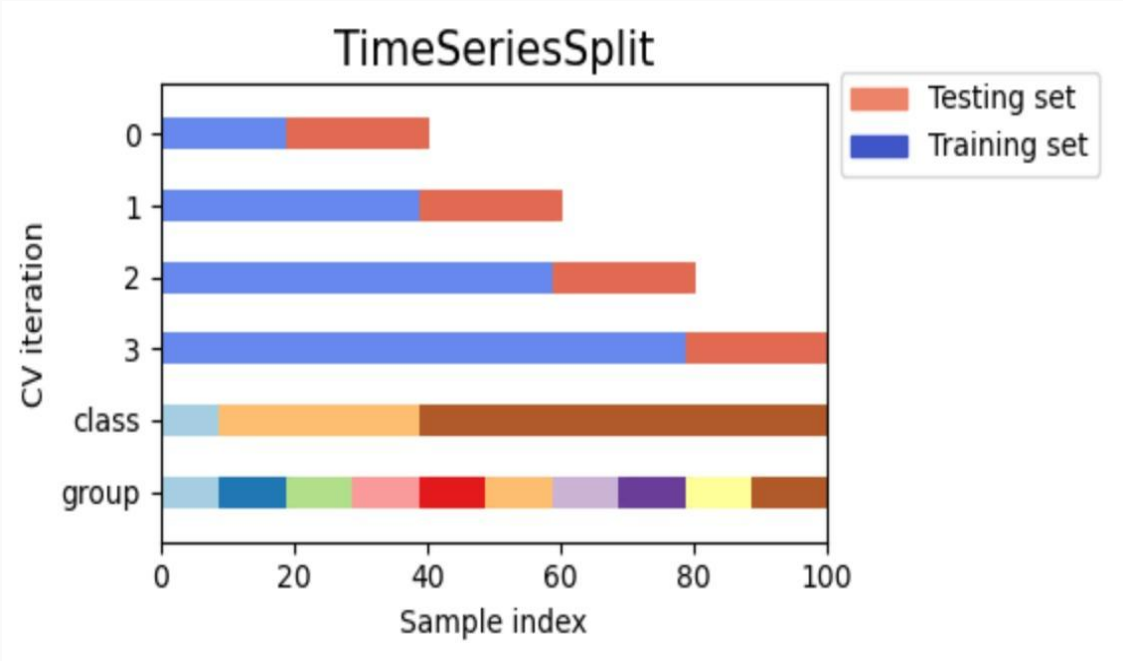


Time Series split

Time Series split is another form of cross validation that can be implemented using the Scikit Learn module

Cross validation using Time Series split breaks the data into smaller parts to test for model robustness over time

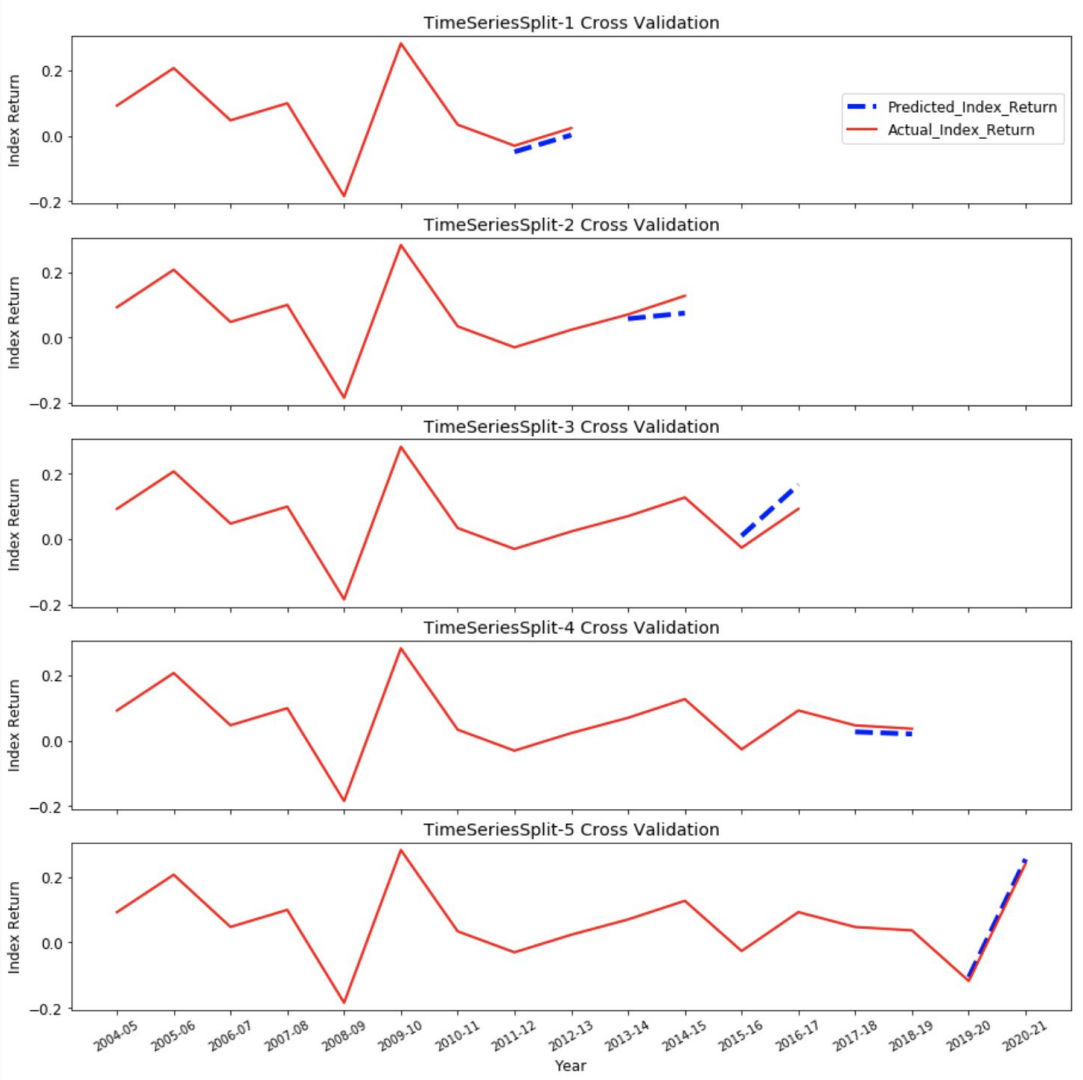
The first set is the smallest, every subsequent set is in addition to the previous set, the final set takes the entire dataset into consideration



Time Series split

Performed a 5 period Time Series split to examine model robustness across different time periods

This form of validation demonstrated that model robustness was highest during the first, fourth and fifth iterations



Time Series split

To better understand the Time Series split validation results I calculated the variance between the actual index returns and the predicted index returns

The deviation between the actual and predicted values ranged from 1% to 7%

The highest deviation occurred for 2016-17 at 7.3% while the lowest was for 2019-20 at 1.2%

```
Year
2011-12    0.018697
2012-13    0.021174
2013-14    0.012952
2014-15    0.052926
2015-16   -0.036410
2016-17   -0.073695
2017-18    0.019856
2018-19    0.016081
2019-20   -0.011943
2020-21   -0.013483
Name: Variance, dtype: float64
```

Conclusion

Some notable learnings during this internship project are as follows -

- Factor investing is a leading trend in the asset management industry, the inflows to factor based investing funds is expected to grow from the current $1.9 trillion to $3.4 trillion in 2022
- R2 and T-stat values are good measures of model robustness but they are not complete measures hence there is need for cross validation
- K-Fold cross validation cannot be used for time series data as it uses future data to predict past performance hence Time Series split and Blocked cross validation techniques are preferred
- For designing a portfolio around a macro factor investing model the ideal range of variance between model results and actual results is 1% - 7 %
- Consistency in magnitude of deviation is important for designing a portfolio, a model that generates substantially accurate predictions for a few years but not for others cannot be used for portfolio implementation

Data sources

- **Nominal GDP** - Data | Ministry of Statistics and Program Implementation | Government Of India
- **Money Supply** - DBIE-RBI > Handbook of Statistics on the Indian Economy > PART II : QUARTERLY/ MONTHLY SERIES > Table 166. Components of Money Stock
- **NEER** - International banking - disseminated data: BIS WebStats
- **Commodity Index** - Historical Data > NIFTY COMMODITIES
- **NIFTY 500 Index** - Historical Data > NIFTY 500
- **CPI** - Thomson Reuters
- **Fiscal Deficit** - Thomson Reuters
- **FII Inflows** - Money Supply - DBIE-RBI > Handbook of Statistics on the Indian Economy > PART II : QUARTERLY/ MONTHLY SERIES > Money and Banking > Table 182. Net Investments by FIIs in the Indian Capital Market
- **10-Year GSec yields** - India 10-Year Bond Historical Data