Trading Signal Exploratory

Team14: Ziqi Shan, Man Shi, Kangjing Shi, Tzuhua(Agnes) Huang

1. Introduction

The purpose of the project is to identify tradable signals in financial markets using funds-flow data. We expect that we can develop an efficient algorithm to predict portfolios' appreciation or depreciation and extract signals from noisy data. The dataset captures the returns and the funds' flow of investments and redemptions within various asset classes (sectors – e.g. technology, finance, consumer goods, etc.)

2. Dataset

2.1 Overview

The dataset contains 20 sectors' data for 3 separate types of investments made in the US.

- Institutional Mutual Fund
- ETF
- Retail Mutual Fund

The weekly data spans 10 years from 2006 through end-Jan 2017. Not all sectors have data available for all the dates since new investment vehicles are introduced at various points in time.

2.2 Data Fields

- **ReportDate**: Weekly data aggregated and released every Wednesday
- **AssetClass**: Industry/Sector/Asset Class
- **Flow**: Amount of positive (inflow) or negative (outflow) in Millions of USD
- **FlowPct**: Flows as percent of assets at beginning of the week
- **AssetsEnd**: Assets at end of the week in Millions of USD
- **PortfolioChangePct**: Percent change in overall portfolio during the week
- **ClosePct**: Close percent change of Russell 2000, SP 500, Dow Jones, and Nasdaq

3. Hypothesis

- ETF has dependency on institutional mutual fund variables and vice versa.
- %Portfolio Change could be a predictive outcome for trading signals.
- Prediction for each sector may be captured by different methods/models.
- Stock market indices can be applied as additional explanatory variables to enhance model predictability.

4. Data Cleaning

4.1 Missing value

First we detect missing values and zeros which would affect the analysis results, we found there is no NA or zero value exist. However, from the statistical description of all the 20 Asset Classes in

three fund types, we know that the ‘Infrastructure’ sector does not have the same number of observations with other sectors. Thus, we also drop ‘Infrastructure’ from our datasets to keep the same dimension of all sectors.

4.2 Outliers

Since we only care about the relationship of tradable signals between Institutional MF and ETF markets in this period, and the percentiles of variables are quite different year by year. Therefore, we handle outliers by replacing them with the 1st and 99th percentile of each variable (FlowPct and PortfolioChangePct) for each year in Institutional MF and ETF datasets. Plot 1 is an example plot of one sector before and after handling outliers. (The red lines represent the original data and black lines show trends after handling outliers)

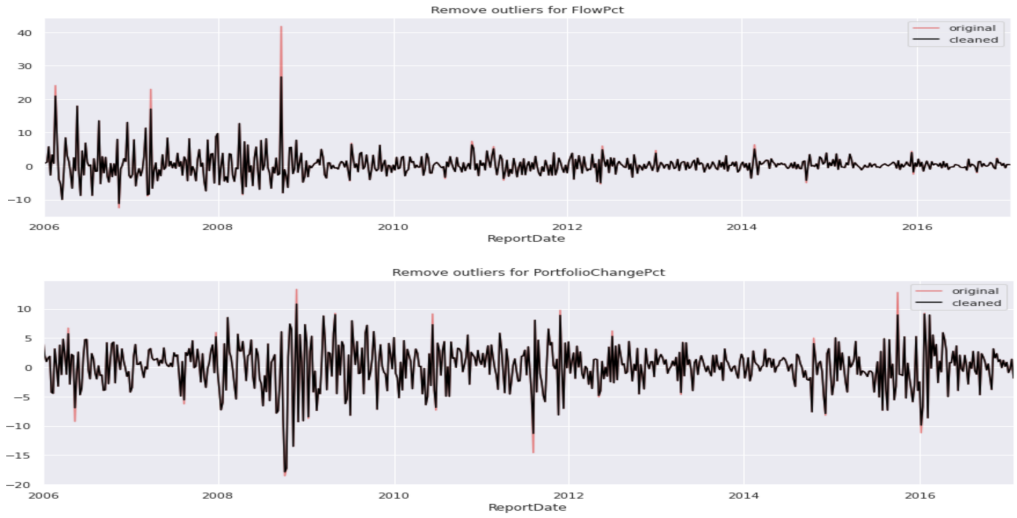

Figure 1. Comparison of data before and after handling outlier

5. Exploratory Data Analysis

In order to have a general understanding of the trend of the markets, we tabulated the average trend of FlowPct and PortfolioChangePct in Institutional MF and ETF in 12 years.

5.1 Average **PortfolioChangePct** for Institution and ETF funds (2006-2017)

From Figure 2. we can notice that the overall trend and range of the mean PortfolioChangePct for ETF and Institutional MF are closely related to each other as well. All sectors showed a huge decrease in 2007-2008 probably resulting from the worldwide financial crisis which caused almost all mutual funds to be sold.

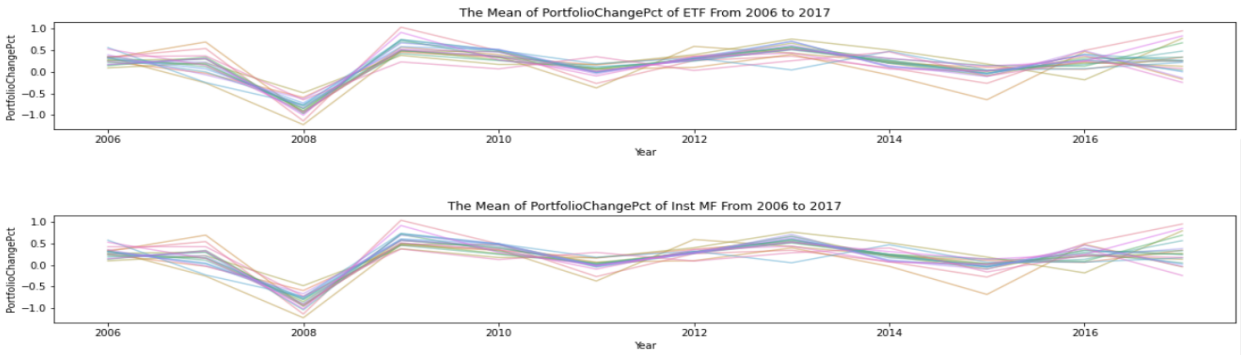

Figure 3. Mean of PortfolioChangePct from 2006 to 2017

6. Time Series Predictive Models

6.1 VAR Model 1 - Cross-Market without Moving Average

Since the key point of our hypothesis is to reveal that the PortfolioChangePct could be predicted by the associated variable FlowPct in different markets, we construct the VAR model both from ETF and Institution to predict tradable signals by identifying actual appreciation or depreciation in Portfolio_etf and Portfolio_ins.

6.1.1 Function

- *%Portfolio_**etf**(t) = a + b1* %Portfolio_**etf**(t-1) + b2* %Flow_**etf**(t-1) + b3* %Portfolio_**ins**(t-1) + b4* %Flow_**ins**(t-1)*
- *%Portfolio_**ins**(t) = a + b1* %Portfolio_**etf**(t-1) + b2* %Flow_**etf**(t-1) + b3* %Portfolio_**ins**(t-1) + b4* %Flow_**ins**(t-1)*

6.1.2 Testing

Before creating the VAR model, we need to construct some tests to verify the validity of the hypotheses in each sector.

- **Granger's Causality Test**
  We first check the causation between variables using Granger's Causality Test, if a given p-value is less than significance level (0.05), then, the corresponding X series (independent variables) causes the Y (response).
- **Augmented Dickey–Fuller (ADF) Test**
  Since the VAR model requires the time series we want to forecast to be stationary, it is customary to check all the time series in the system for stationarity. If a series is found to be non-stationary (p-value > 0.05), we can make it stationary by differencing the series once and repeat the test again until it becomes stationary.
- **Durbin Watson Statistic Test**

  If there is any correlation left in the residuals, there is some pattern in the time series that is still left to be explained by the model. The value of this statistic can vary between 0 and 4.

The closer it is to the value 2, then there is no significant serial correlation. The closer to 0, there is a positive serial correlation, and the closer it is to 4 implies negative serial correlation. We have a majority of values around 2, suggesting there is no autocorrelation detected in the residuals.

6.1.3 Fit the model and predict %Portfolio

We divide both the Institutional MF and ETF into train and test sets, the test sets include last 10 weeks data (from 2016-11-30 to 2017-02-01) and the rest are train sets. Then we fit the VAR model with train sets and predict 10 values and compare with the test sets. Figure 4. shows an example of the actuals versus forecasts (Energy sector).

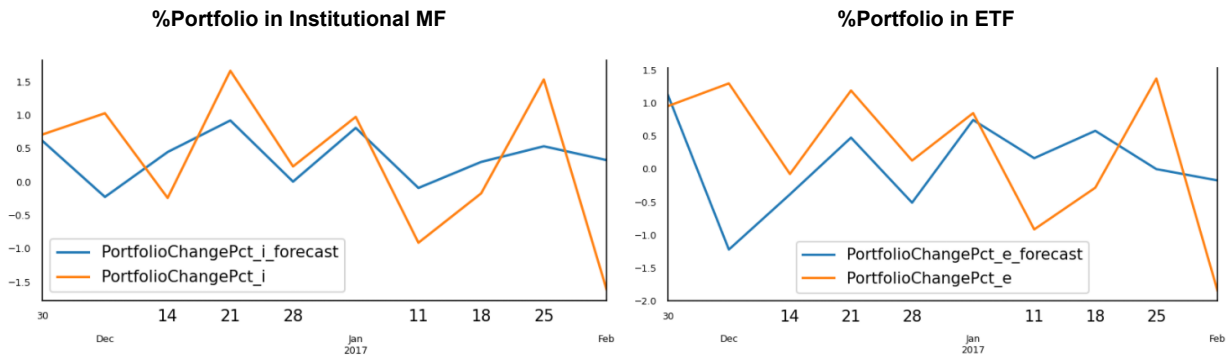

Figure 4. Actuals v.s. Forecasts of Energy Sector

Next, to better understand the directional patterns of the forecasts, we change actual and forecast values to binary numbers to detect direction. We set the threshold as 0.1 which means if the value < last value - 0.1, then change the value to -1; if the value > last value + 0.1, then change the value to 1; otherwise change the value to 0. (The threshold 0.1 was computed by taking the average of the difference between this week data (t) and last week data (t-1) in the 10 actuals.) Figure 5 shows an example of the actuals versus forecasts after changing to directional values (Energy sector).

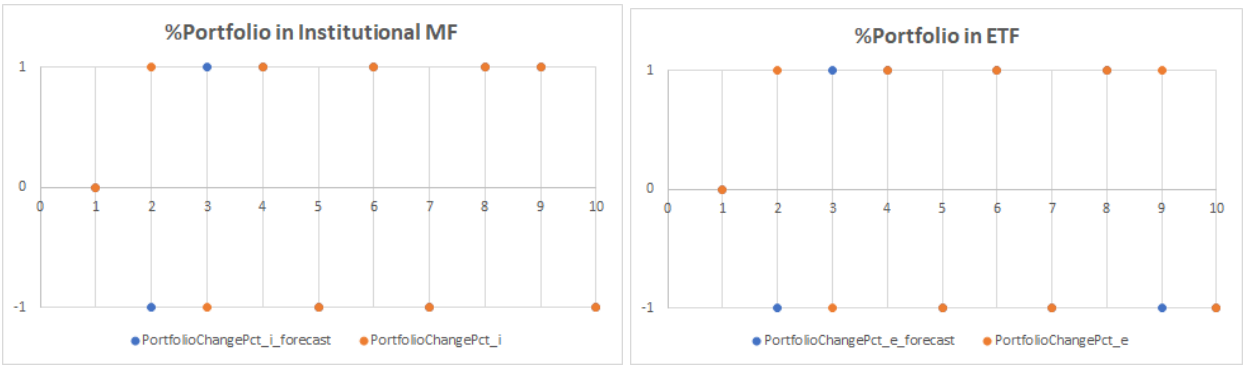

Figure 5. Actuals v.s. Forecasts (directional values) of Energy Sector

6.2 VAR Model 2 - Cross-Market with Moving Average

In this part, we would like to construct the VAR model in cross-market and apply the Moving Average smoothing method.

6.2.1 Function

- *%Portfolio_**etf**(t) = a + b1* %Portfolio_**etf**(t-1) + b2* %Flow_**etf**(t-1) + b3* %Portfolio_**ins**(t-1) + b4* %Flow_**ins**(t-1)*
- *%Portfolio_**ins**(t) = a + b1* %Portfolio_**etf**(t-1) + b2* %Flow_**etf**(t-1) + b3* %Portfolio_**ins**(t-1) + b4* %Flow_**ins**(t-1)*

6.2.2 Testing

We use the same tests as previous models to verify the validity of the hypotheses in each sector. Besides, we apply the moving average method trying to improve the accuracy of forecasts. Since one month could be regarded as a feasible fund market period, we choose the windows=4 (one month) as the parameter for moving average smoothing. It returns us new datasets which are smoother and more useful in the prediction process. Figure 6 shows the comparison of FlowPct and PortfolioChangePct before and after moving average (Energy sector in ETF).

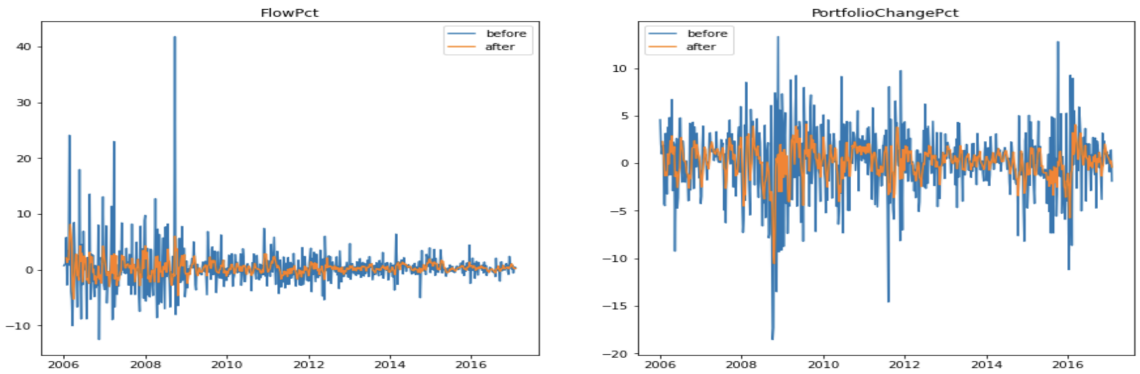

Figure 6. Data comparison before and after moving average (Energy in ETF)

6.2.3 Fit the model and predict %Portfolio

We divide both the Institutional MF and ETF into train and test sets, the test sets include last 10 weeks data (from 2016-11-30 to 2017-02-01) and the rest are train sets. Then we fit the VAR model with train sets and predict 10 values and compare with the test sets. Figure 7 shows an example of the actuals versus forecasts (Energy sector).

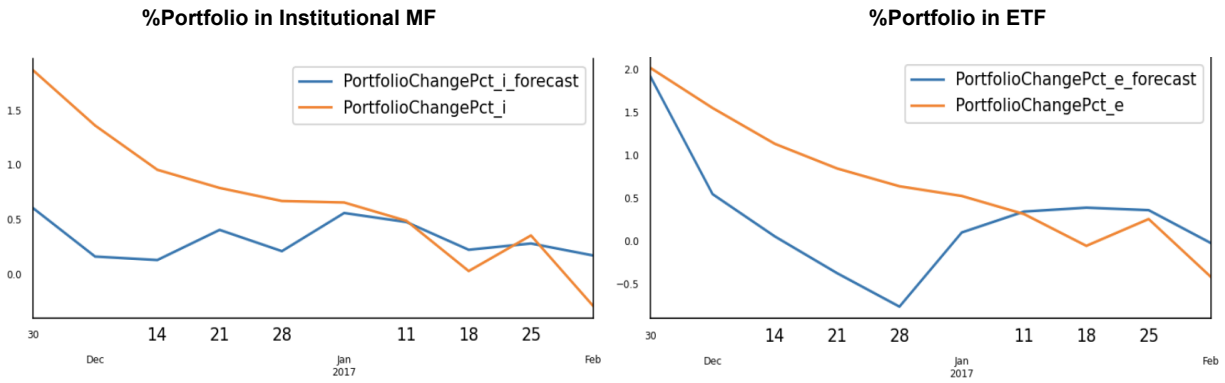

Figure 7. Actuals v.s. Forecasts of Energy Sector

Next, we also change actual and forecast values to binary numbers to detect direction. Figure 8 shows an example of the actuals versus forecasts after changing to directional values (Energy sector).

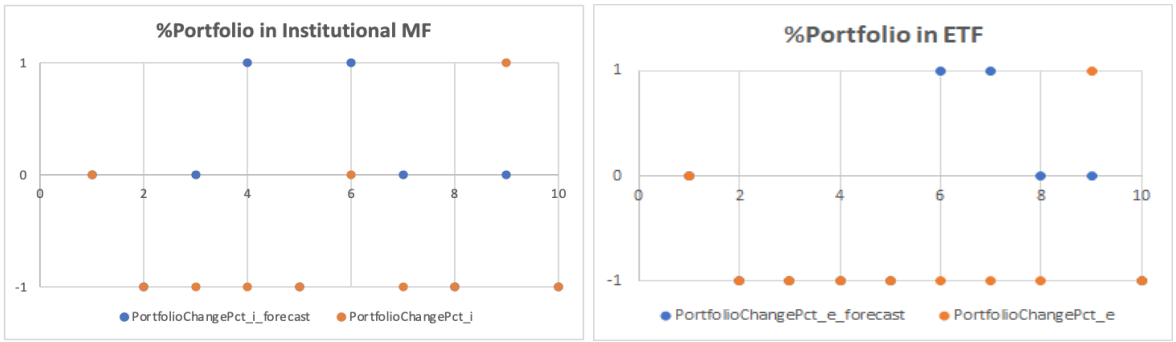

Figure 8. Actuals v.s. Forecasts (directional values) of Energy Sector

6.3 VAR Model 3 - Cross-Market with Moving Average, market indices applied

After the cross-market VAR models, we would like to find a way to further improve our predictions. Therefore we add new market indices variables, trying to capture the influence of markets to our classes' Portfolio value.

6.3.1 Function

- *%Portfolio_**etf**(t) = a + b1* %Portfolio_**etf**(t-1) + b2* %Flow_**etf**(t-1) + b3* %Portfolio_**ins**(t-1) + b4* %Flow_**ins**(t-1) + **V***
- *%Portfolio_**ins**(t) = a + b1* %Portfolio_**etf**(t-1) + b2* %Flow_**etf**(t-1) + b3* %Portfolio_**ins**(t-1) + b4* %Flow_**ins**(t-1) + **V***
  ***V** = related market index (SP 500, Dow Jones, Nasdaq, Russell 200) ClosePct*

6.2.2 Allocation

We did not use all four market indices for every class. Instead, we allocate the proper indices to each class based on the class type.

	SP 500	Dow Jones	Nasdaq	Russell 2000
Commodities		V		
Consumer Goods	V	V		
Energy		V		
Financials	V	V		
Health Care/Biotech	V			
Industrials		V		
Large Cap Blend	V			
Large Cap Growth	V			
Large Cap Value	V			

	SP 500	Dow Jones	Nasdaq	Russell 2000
Mid Cap Blend				V
Mid Cap Growth				V
Mid Cap Value				V
Real Estate	V	V		
Small Cap Blend				V
Small Cap Growth				V
Small Cap Value				V
Technology			V	
Telecom		V		
Utilities		V		

6.2.3 Fit the model and predict %Portfolio

We divide both the Institutional MF and ETF into train and test sets, the test sets include last 10 weeks data (from 2016-11-30 to 2017-02-01) and the rest are train sets. Then we fit the VAR model with train sets and predict 10 values and compare with the test sets. Figure 9 shows an example of the actuals versus forecasts (Energy sector).

%Portfolio in Institutional MF **%Portfolio in ETF**

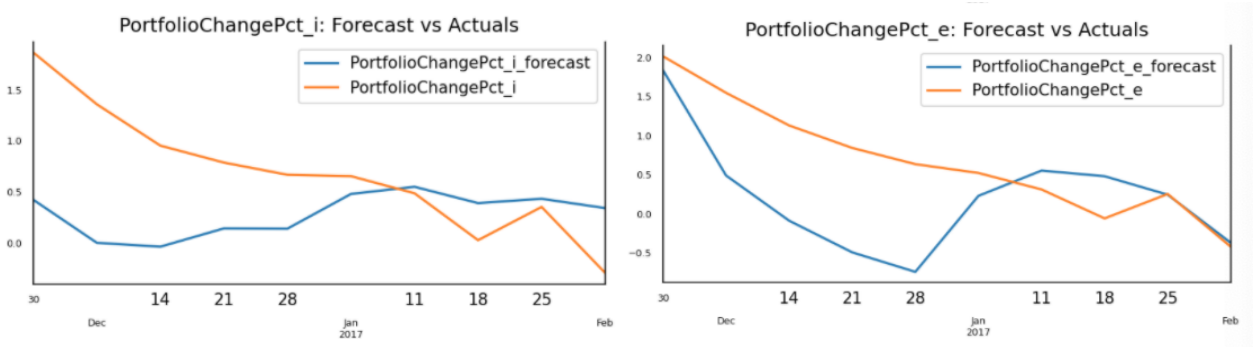

Figure 9. Actuals v.s. Forecasts of Energy Sector

Next, we also change actual and forecast values to binary numbers to detect direction. Figure 10 shows an example of the actuals versus forecasts after changing to directional values.

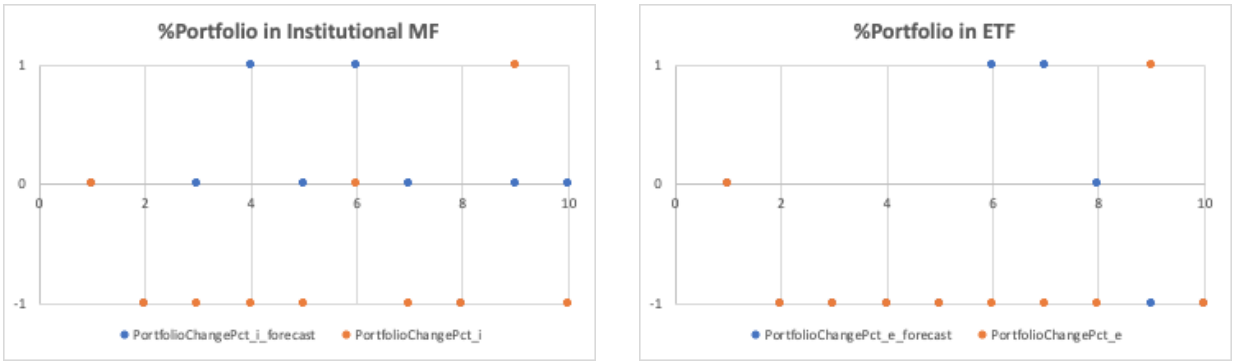

Figure 10. Actuals v.s. Forecasts (directional values) of Energy Sector

6.4 LSTM (Long Short-Term Memory) Network

LSTM, an advanced RNN, is able to store past information and learn long-term order dependencies; it is powerful in time sequence prediction problems. According to our hypothesis, we assume that the previous portfolio change of a sector is crucial in predicting its future change.

6.4.1 Process

6.4.1.1 Data Preparation

For illustration, we use a subset of data from the energy sector as a pilot data set. For demonstration below, we will predict the portfolio change of institutional mutual funds.

a. **Splitting Data**
   In this part, we fit an LSTM on the univariate input data. First, we split the prepared dataset into train and test sets. In this part, we fit the model on the first 9 years of data, then evaluate it on the remaining 1 year of data.
b. **Scaling and Transforming Data**
   To help the LSTM model to converge faster, we scale our data for optimal performance. In our case, we use 'MinMaxScaler' in 'scikit-learn' and scale our dataset to numbers between zero and one. After scaling, we need to transform the data into a format that is appropriate for modeling with LSTM.
c. **Reshaping Data**
   LSTM expects the input data to be in a specific format, we reshape the inputs (X) into a 3-dimension format, namely [samples, timesteps, features]. We start by creating data in 52 timesteps (i.e. one year period) and converting it into an array using NumPy. Next, we convert the data into a 3-dimensional array with X_train samples, 52 timesteps, and one feature at each step.

6.4.1.2 Define and fit the model

a. **Required modules for LSTM**
   To build the LSTM, we need to import modules from Keras as follows:
   - Sequential: initialize the neural network with a plain stack of layers
   - Dense: deeply connected neural network layer
   - LSTM: Long Short-Term Memory layer
b. **Building the LSTM**
   I. We define the LSTM with 50 neurons in the first hidden layer and 1 neuron in the output layer for prediction. The input shape as mentioned above, which is 52 timesteps with 1 feature.
   II. We set the Mean Absolute Error (MAE) as loss function and use the Adam as an optimizer to deal with stochastic gradient descent.
   III. The model fit for 100 training epochs with a batch size of 32. The internal state of the LSTM in Keras is reset at the end of each batch, so an internal state that is a function of a number of days may be helpful.
   IV. Finally, we track both the training and test loss during training by setting the validation_data argument in the fit() function. At the end of the operation, we plot both the training and test losses.

6.4.1.3 Evaluate Model

a. **Forecasting**
After the model is fit, we can forecast on the test dataset. We combine the forecast with the test dataset and use inverse_transform to invert the scaled response variables to the original readable format. Then, we use Matplotlib to visualize the result of the actual and the predicted PortfolioChangePct. Additionally, we calculate the Root Mean Squared Error (RMSE) that gives the error between two sequences of values.

b. **Converting to binary variables**
With the forecasts and the actuals valued in the original scale, we convert the results from numbers to binary numbers as we did within the VAR model. To detect the change in direction, we have to define a proper threshold when we make conversion. We loop over a range of thresholds and determine the best performing threshold by choosing the one with the highest accuracy score.

c. **Comparing the Actuals v.s. Forecasts**
To visualize the comparison between actual and predicted change in portfolio returns, we plot converted values in the same figure to observe if there is significant overlap. In addition, to better understand and quantify the accuracy of prediction, we will calculate the accuracy score of each model as demonstrated in Table 1(below).

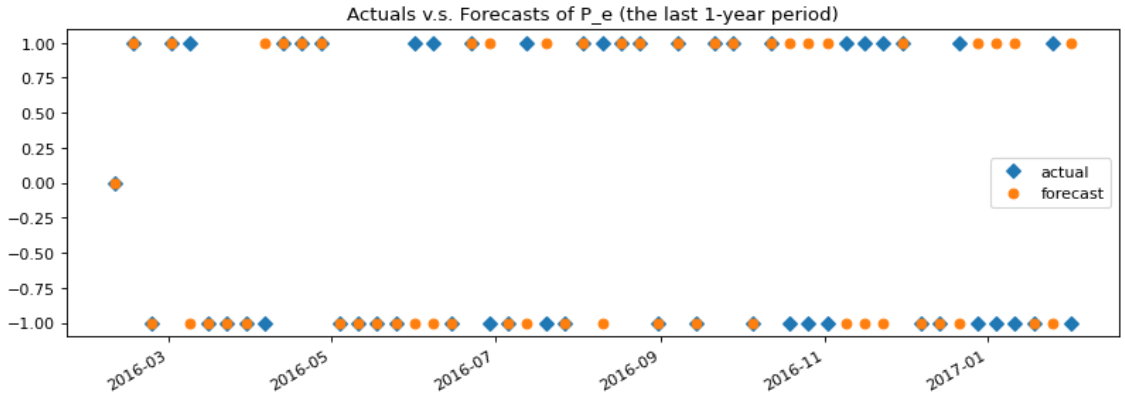

Figure 11. Actuals v.s. Forecasts (directional values) of Energy Sector

6.4.2 Summary

Since our objective is to detect tradable signals in the financial market, we define the criteria as the best-performing threshold with the highest accuracy score by each sector. In the last step, we iterate the above workflow and evaluate the performance of the models by calculating the accuracy score for each sector, and we summarize the results in the table below.

	Consumer Goods	Commodities	Energy	Finance	Health/ Biotech	Industrials	Large Capital Blend
P_e	0.54	0.40	0.62	0.65	0.63	0.56	0.50
P_i	0.54	0.44	0.63	0.65	0.65	0.56	0.62

	Large Capital Growth	Large Capital Value	Mid Capital Blend	Mid Capital Growth	Mid Capital Value	Real Estate	Small Capital Blend
P_e	0.58	0.44	0.54	0.42	0.65	0.44	0.46
P_i	0.50	0.50	0.56	0.56	0.56	0.50	0.50
	Small Capital Growth	Small Capital Value	Technology	Telecom	Utility		
P_e	0.60	0.54	0.46	0.50	0.58		
P_i	0.62	0.56	0.52	0.52	0.62		
● P_e: %Portofolio Change in ETF ● P_i: %Portofolio Change in Institution Mutual Fund							

Table 1. Accuracy Score of Predictive Models in Cross Market (MA=4) Applied LSTM Network

7. Conclusion

a. Compare VAR model and LSTM

Comparing VAR models with the LSTM model, we found that the simple model(VAR) performs better than the complex one(LSTM). According to table1 above, for the LSTM model, the accuracy score of each sector is all around 0.5/0.6. However, the VAR model could give us a better result for several classes with an accuracy score up to 0.9.

b. Accuracy Score with Applying VAR Model

According to our hypothesis, we assume that portfolio change of the ETF and institutional mutual funds are interdependent. We will compare the model efficiency based on the model applying cross-market variables. Considering the performance of 3 models, we summarize the accuracy scores as the table below.

		Consumer Goods	Commodities	Energy	Finance	Health/ Biotech	Industrials	Large Capital Blend
M1	P_e	0.6	0.4	0.7	0.4	0.7	0.5	0.6
	P_i	0.2	0.7	0.8	0.5	0.5	0.5	0.7
M2	P_e	0.3	0.6	0.6	0.4	0.9	0.5	0.5
	P_i	0.2	0.4	0.5	0.7	0.4	0.6	0.3
M3	P_e	0.4	0.3	0.6	0.3	0.8	0.7	0.6
	P_i	0.5	0.4	0.3	0.6	0.5	0.6	0.5
		Large Capital Growth	Large Capital Value	Mid Capital Blend	Mid Capital Growth	Mid Capital Value	Real Estate	Small Capital Blend
M1	P_e	0.3	0.5	0.4	0.6	0.5	0.8	0.4
	P_i	0.5	0.5	0.3	0.3	0.3	0.3	0.5
M2	P_e	0.6	0.4	0.7	0.5	0.6	0.4	0.5

	P_i	0.4	0.4	0.5	0.3	0.5	0.3	0.6
M3	P_e	0.6	0.4	0.4	0.4	0.8	0.4	0.7
	P_i	0.6	0.4	0.6	0.3	0.5	0.3	0.7
		Small Capital Growth	**Small Capital Value**	**Technology**	**Telecom**	**Utility**		
M1	P_e	0.6	0.4	0.6	0.6	0.5		
	P_i	0.2	0.4	0.5	0.7	0.5		
M2	P_e	0.5	0.7	0.2	0.4	0.5		
	P_i	0.5	0.4	0.3	0.3	0.6		
M3	P_e	0.6	0.7	0.4	0.4	0.4		
	P_i	0.5	0.4	0.2	0.3	0.6		
● P_e: %Portofolio Change in ETF ● P_i: %Portofolio Change in Institution Mutual Fund								
● M1: VAR Model 1 - Cross-Market without Moving Average ● M2: VAR Model 2 - Cross-Market with Moving Average ● M3:VAR Model 3 - Cross-Market with Moving Average, market indices applied								

c. Identify Tradable Signal

To explain the tradable signal, we take one of the best performing models for example. As shown in the plot below, we can see that some of the actual and predicted values overlap. Based on the criteria we defined,we classify the predictions into three classes: 1 (increase in portfolio change), -1 (decrease in portfolio change), and 0 (non-change). In this case, if the model generates a desirable accuracy score, the overlapped dots indicate that forecasts capture actuals. For example, when we choose a specific timestamp with both actual and forecast located in the class labeled 1, we can infer that the model successfully predicts the positive movement of portfolio change, and view it as a signal to invest in the sector. According to the model's accuracy score, which is 0.8, we can conclude that approximately 80% of the portfolio's movements are captured by the model.

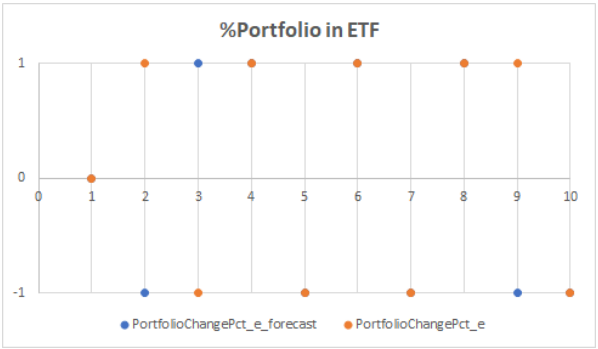

Figure 12. Actuals v.s. Forecasts (directional values) of Energy Sector