Capstone Project Final Report

Team A6

Kaiyu Wang, Simeng Li, Xiangshan Mu, Yihan Jia, Yinghao Wang

I. Code Repository Location:

For our capstone project, we saved our codings into our member's public GitHub repository:
Link: [Capstone Team A6 Github Repository]

II. List of Tools & Documentation Developed:

Throughout the whole process of our capstone project, all the coding work is done via Google Colab with Python programming language. The Python packages we have used are listed as follows:

- pandas, numpy, and datetime for basic data import, statistical description, and pre-processing;
- matplotlib and seaborn for plotting;
- statsmodels for the first 3 forecasting models of the time series analysis;
- pytorch for long short-term memory estimations from recurrent neural network architecture;
- scikit-learn for multi-class classification forecasting.

Our team developed the following documentation in use:

1. Midterm & Final presentation slides;
2. Poster presentation;
3. Integrated coding report;
4. Final report and summary

III. Key Findings and Deliverables:

About Dataset:

The dataset provided by our BA includes 3 different types of funds conducted in the US and contains 60-70% of the activity in the overall asset class. Each row of observations is weekly recorded and spans a 10-year period from 2006 to the end of January 2017.

About the Dataset:

The dataset contains sectoral data for 3 separate types of investments made in the US. It represents 60-70% of activity in the overall asset classes for that week. The weekly data spans 10 years from 2006 through end-Jan 2017.

- **Institutional Mutual Fund Holdings (IMF)** - For institutional
- **Retail Mutual Fund Holdings (RMF)** - For individual
- **Exchange Traded Funds (ETF)** - The convenience of trading is available at any time during the day.

Data Fields Description:

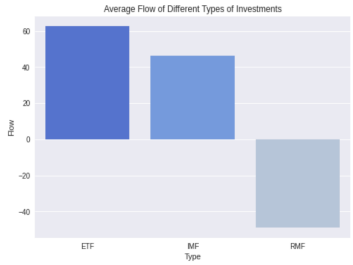
- **ReportDate**: Weekly data aggregated and released every Wednesday
- **AssetClass**: Industry/Sector/Asset Class
  - Industry: 20 industries for each type of fund (3 types of funds)
  - Sector: North America-USA-North America (one value for all)
  - Asset Class: Equity (one value for all)
- **Flow**: Amount of positive (inflow) or negative (outflow) in Millions of USD
  - As the label of time series
- **FlowPct**: Flows as a percent of assets at beginning of the week
- **AssetsEnd**: Assets at end of the week in Millions of USD
  - It includes factors outside the dataset that cannot be measured, so it can not be used as a label
- **PortfolioChangePct**: Percent change in the overall portfolio during the week

Data Basic Preprocessing:

We first merged the three data sets into one data frame for further analysis and added a "type" feature to identify the categories. In the process of cleaning the data, we detected no null values and no duplicate rows. However, we chose to drop all asset classes and

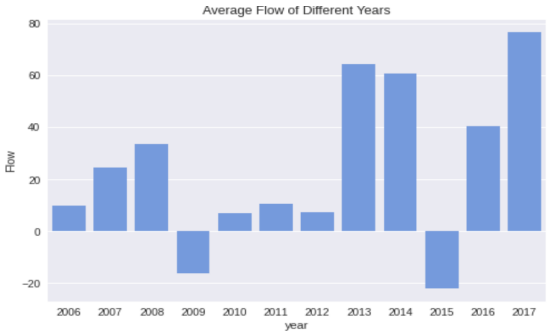
sectors as they were shown as the same. Finally, we extracted the year and month from the dates for further EDA and classification modeling.

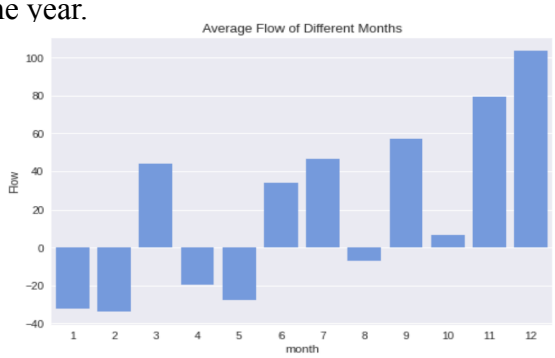
Exploratory Data Analysis:



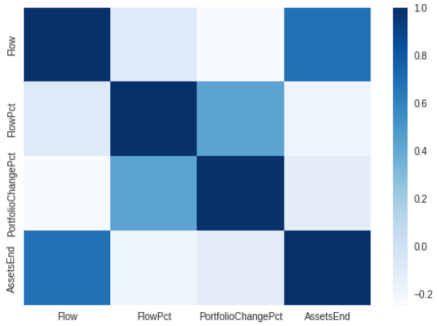
Next we analyzed the feature 'Flow' and compared the average flow of different types of funds. As shown on the right, we found that ETF type has the largest average flow and RMF has the smallest and negative values. We also studied the average Flow of different years and months. From the plot below we see that the average flows from 2013 to 2017 were very high, except for 2015.

And average flows in 2009 and 2015 were negative, and we would guess more funds were withdrawn at that time. From the plot on the right below, we can see that most of the positive flow is concentrated in the second half of the year. We could speculate that most of the funds flowed in in the second half of the year.

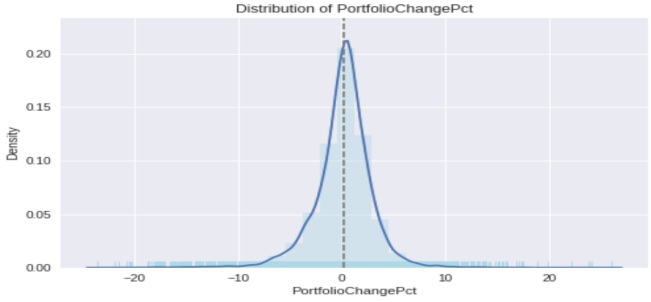


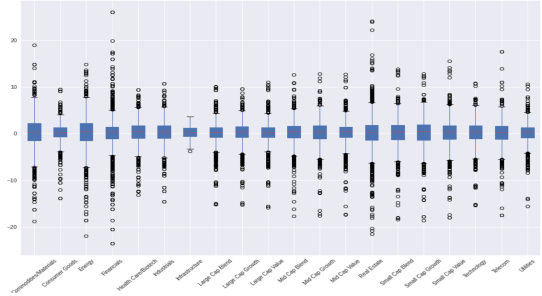


Then we compared the correlation between different features. From the heatmap, we know more clearly that the correlation between Flow and AssetsEnd is large, but the correlation between Flow and PortfolioChange-Pct is extremely small, so next step we will focus more on the feature PortfolioChangePct.



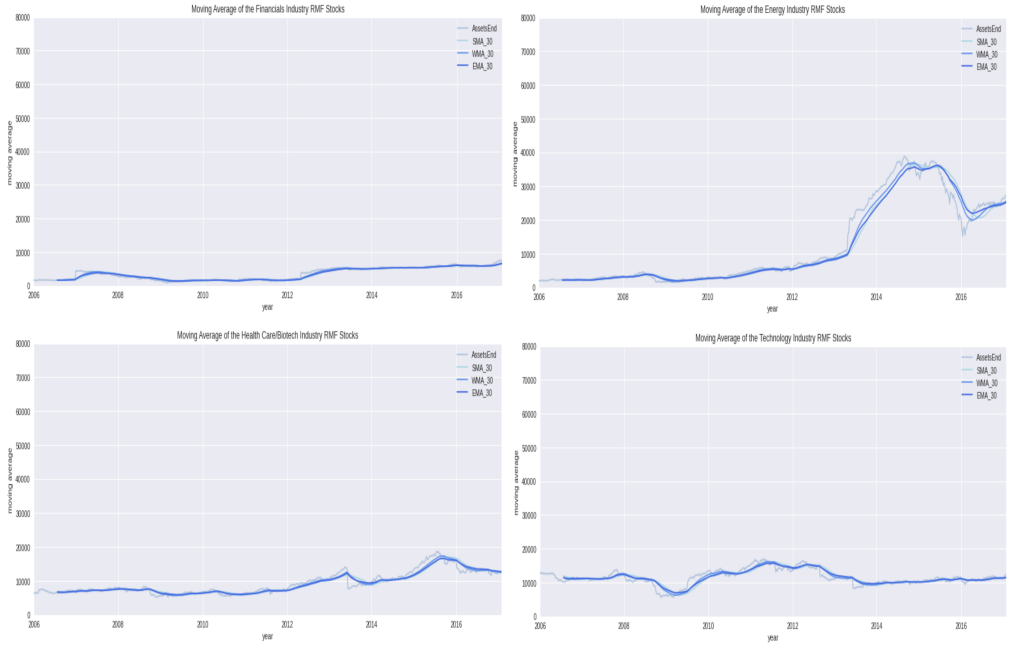
As we can see from the plot on the left below, the feat-are PortfolioChangePct is normally distributed, The graph on the right shows the average PortfolioChangePct for different industries. The Financial and Energy industries have the smallest average values, which partly means that these 2 industries are more stable and more suitable for cautious investors. Conversely, Healthcare and Technology industries are riskier with the largest average values, and more suitable for open investors. Based on the result of the average PortfolioChangePct, we started to focus on the 4 most representative industries: Financial, Energy, Health Care and Technology.





The distribution of PortfolioChangePct as shown above shows the peaks and valleys that each industry has ever reached. Combining the previous result, we know that the Financial & Energy industries have small fluctuations and large distribution, which means that it is easier for investors to make profits, cause they may have enough time to consider selling at the peak. Healthcare and Technology are highly volatile but in small distribution, which means that even a drop won't result in huge losses.
Also, we applied 3 moving average calculation methods to smooth the data. And we set the window size to 30 to make the line smoother. The following plots show the AssetsEnd trend of 4 industry ETF funds. Also, AssetsEnd of IMF funds moves in almost the same way as that of ETF funds, but volatility is a little bit greater than ETF.

Besides, the AssetsEnd value of RMF funds is the smallest and smoothest. From a forecasting perspective, energy companies have the highest values of AssetsEnd.



See from different types of funds, we found they have different natures. AssetsEnd of IMF funds and ETF funds move almost the same, but the IMF is more volatile, but there is no big difference between IMF fund annual returns and ETF returns. RMF-type funds have the smallest and smoothest AssetsEnd values; funds in the same industry tend to have a similar trend and range in terms of Flow for IMF and ETF types.

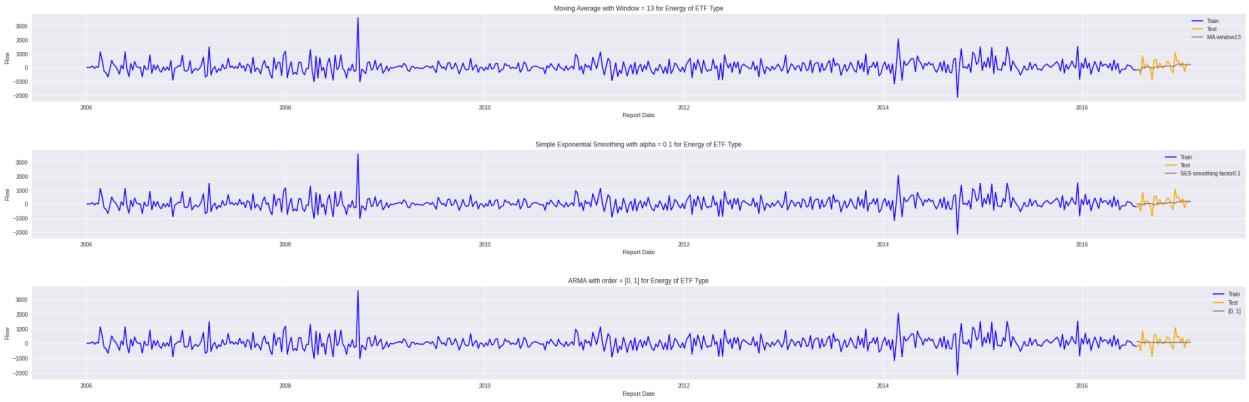
Time Series Analysis:

For our local optimization approach to predict future flow value with multiple time series forecasting models, we generate 4 different models for each unique combination of industry and fund type for short-term forecasting: moving average, simple exponential smoothing, autoregressive moving average, and long short-term memory.

The first step is to take the stationary check for each input subset. Only when we ensure that all the subsets of flows are proved to be stationary, which means there are no trend, seasonality, or cyclical pattern shown, we can then apply selective models to the dataset.

In terms of general design for models, the only independent variable here is time, and the dependent variable is our target signal flow value. For validation purposes, we split each subset into train and test sets. By doing this we feed past time series data as input and use the fitted one to predict the future flow. The evaluation metric we used here is the mean squared error, and the smaller the mean squared error is, the better result our model returns.

We iterate the model parameters for each subset of flow to figure out the best-performed one when we apply the first three model options. Take one Energy industry from ETF type group as an example, we plotted the predicted values and actual flow in the following plots:



We can observe that the predicted line is smooth, however not exhibit a perfect accuracy in terms of the exact value. Also, the result for moving average methods are all with greater windows, while for simple exponential smoothing are with smaller smoothing levels and small autoregressive factors. This illustrates our original flow of ten years' history behaving relatively stable within a certain range.

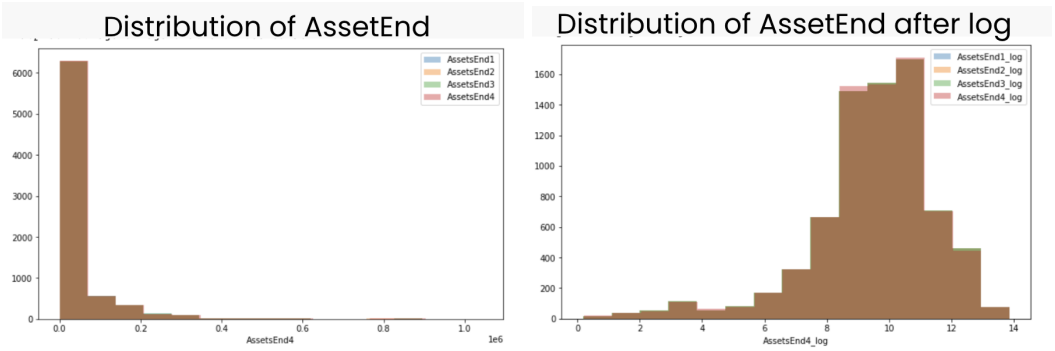
As presented in chart III-1, long short-term memory with the help of standard scaled input returns an almost perfect forecasting result, which means the model not only follows the ups and downs of value change but also does well in predicting exact values. Therefore, we can see that while time series analysis offers a simple and efficient way for short-term forecasts for a specific fund, it can also generate a satisfactory output.

Multi-Class Classification:

Multi-Class Classification is our second method to find the tradable signal. The concise workflow (chart III-2) is that we first combined the three datasets as the original data. Later on, we did data processing and features engineering. Then we fit the processed data into 10 classifier models, still, Gradient Boost Decision Tree is the best classifier. The main idea of our global prediction is to predict which interval the next month's Flow will locate in, based on the indicators of this month. One of our innovations is in our self-defined tradable signals. There are 3 tradable signals which are: sell, hold and buy. For example, when the classifier predicts the next month's flow will be less than minus 100, then the recommendation is to sell. The same as the hold (the next month's flow is located in (-100,100]) and buy (the next month's flow is located in (100,]).

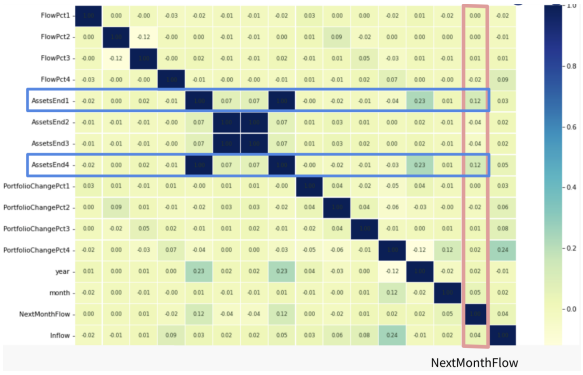
For the data processing, we did 3 things to make our model perform better: First, we dropped the infrastructure industry, because it only has 67 observations which are too limited to study, and might be treated as noise in our model. Second, we dropped not fully observed months, because later on, we need to group weekly data into monthly, the

not fully observed data could not be integrated into one line. Third, we used the winsorization method to process the outlier. Also from the 2 plots at the bottom, you can see the distribution of the asset end on the left is right skewed, so we log the asset end to make it normal distributed, which will contribute to the training.



For the features engineering, We have integrated 4 weeks of data as one observation because there would be more predictors in one line which would make our model more reliable and perform better. Additionally, we created more features. For example, we added a bool feature named inflow, “1” means the flow in this month is a positive number, and “0” means negative. Furthering, we did normalization and one hot encoding on the features. Last, we bucketed the continuous data of our tradable signal (next month flow) into categorical data which are “sell”, “hold” and “buy” in three categories. The final size of the processed data is 7493 rows × 11 columns. A sampling of the data is 80% train and 20% test.

After data processing and features engineering, we print out the correlation heatmap, we focused on the coefficients in red which is predictors of correlation to our tradable signal, the next month's flow. The most powerful indicators are the first and fourth week’s AssetEnd. The processed data are more powerful in predicting the tradable signal but still cannot generate enough correlation.



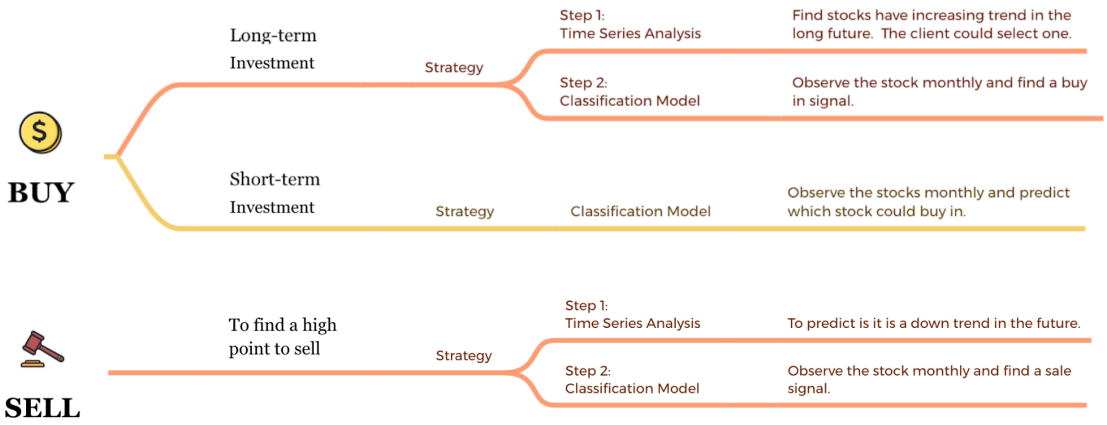
We have fitted the data into 10 classifiers and printed out their accuracy, precision, recall, and confusion matrix. Compared with our 10 classifiers, GBDT has the best performance. The overall accuracy is at the level of 67.2% The precision for tradable signals (sell, hold, buy) is 64%, 74%, and 64% respectively.

```
Start Training: GBDT

accuracy is 0.672448

              precision    recall  f1-score   support

          -1       0.64      0.65      0.65       507
           0       0.74      0.69      0.71       459
           1       0.64      0.68      0.66       533

    accuracy                           0.67      1499
   macro avg       0.68      0.67      0.67      1499
weighted avg       0.67      0.67      0.67      1499

Confusion Matrix...
[[331  46 130]
 [ 74 315  70]
 [109  62 362]]
```

One advantage of our classification is we can give a global optimization using only one model. However, there are also some limitations in our model building. First, we did not consider the currency value has changed during the past 10 years. Changing the trading signal as ratios might be a solution. Second, more features are needed to show the performance of the stock, from which the model can capture more information.

IV. Business Impact and Applications:

The prediction of fund prices is of great significance in business and finance. Stock traders, mutual funds, investors, and security analysts all want a way to predict fund price fluctuations. Anticipating the fluctuations of stock prices and tradable signals in advance can help fund traders make better investment decisions. We can also make recommendations based on long-term, short-term, and different client needs and fund types. For example, if the flow is predicted to increase and potentially reach the peak in the next two weeks, we can suggest clients sell it at that time for more profits.
Also, we provide the following workflow chart as the summary use case scenario of our project:



V. Limitations and Avenues for Future Improvement:

Although our two methods (time series analysis and multi-classification) can identify the tradable signals, the prediction is not very precise and reliable. A further step is to do data mining, such as searching for financial data of funds to be added to our training data to improve the performance and reliability of our model. We have applied time series models and multi-classification models which are too complicated to interpret. One of the limitations is we do not know how the model figures out the tradable signal. We do not know whether the model's solution is reasonable for the commercial experience. In order to get a better understanding of the model and data, we should also use a simple model like linear regression.

From our time series analysis, the first three models (MA, SES, and ARMA) include smoothing factors or averaging methods that offset the ups and downs of flow fluctuations, which weakened the accuracy of predicting the true flow value to a great extent. We can try to reduce the size of the smoothing window and find a balance point that can detect fluctuations in the data and make the data more stable. Even if the LSTM model from the recurrent neural network architecture could follow the most jump and down and return with a predicted flow value that is very close to the actual one, we still do not know what happened inside of the 'black boxes' of neural network hidden layers. We can try to find a more interpretable model to explain the tradeable signals. For the classification model, the accuracy for this model is around seventy percent, we could try to do more feature engineering and data processing to improve the accuracy.

VI. One-Page Summary:

Through the early exploratory analysis, we dived into the dataset and found more information about feature characteristics. We can figure out that different industries vary widely in terms of the flow range, change percentage, and overall trend in ten years. We can also figure out that for funds with the same industry but different types, usually the ETF type and RMF type share the same pattern. One key finding from this part is that we can distinguish funds based on their volatility and stability. For example, Financial & Energy types of funds are more stable inflow value and easier to gain profit but the annual rate of return shows a downward trend instead. While for Health Care/Biotech and Technology, their volatility is high and thus is riskier for the possibility of taking huge losses, but the annual rate of return is on an uptrend. These results can be used for clients to generate an overview of funds' risk rate at a very first glance.

We need to use machine learning applications to understand the underlying logic of the data and also to figure out how to detect the tradable signals, and also give reasonable suggestions for the fund trading market clients and managers. We have used two methods to find the tradable signals. One is time series analysis with a local optimization approach, the other one is a global optimization method with multi-classification models.

Our time series analysis can be used to distinguish the stability and volatility of funds, while also giving a short-term prediction based on individual fund types and industry. The first three models are with formulas that are intuitive to be understood, while the involved neural network model returns a high accuracy which can tell exact future flow value. Our classification method could help clients predict the specific flow range for the future month, while also forecasting the trend of fund market fluctuation. Therefore, with our two approaches combined together, we can give personalized suggestions on fund trading based on current market situation and clients' on-hand portfolios.

Detecting the tradable signal and generating an accurate prediction of fund prices is of great significance in business and finance. Anticipating the fluctuations of fund prices in advance can help fund traders make better investment decisions. We can also make recommendations based on long-term, short-term, and different client needs and fund types. For example, if the overall trend of a fund that the client wants to buy is down in the next few months, we can suggest they wait instead of buying immediately.

One thing that needs to be noticed is that our project still has flaws and deficiencies. Even if our time series analysis can generate trend forecasting, funds' best models are always with a large window that can offset flow's ups and downs. Also, since the observations' time span is ten years long, in current fund situation which is largely influenced by factors outside the market, the ability to predict is relatively low.

VII. Appendix:

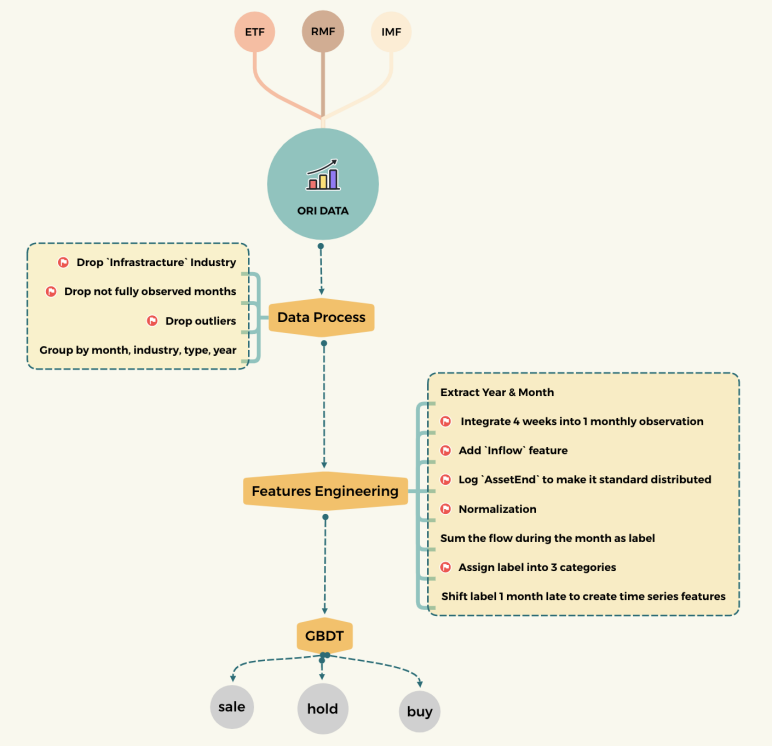
Code Section Locations:
To reduce the rendering time and for better performance, we split the integrated Jupiter notebook into the following sections:

1. Data field descriptions and basic preprocessing (Section 01)
2. Exploratory data analysis and statistical analysis (Section 02)
3. Time series forecasting (MA vs SES) (Section 03)
4. Time series forecasting (ARMA) (Section 04)
5. Time series forecasting (LSTM approach) (Section 05)
6. Multi-class classification (Section 06)

Chart III - 1:

Model	MAE	MSE
MA(13)	410.30	1.683463e+05
ES(alpha = 0.1)	425.51	1.810559e+05
ARMA(0, 1)	408.61	1.669611e+05
LSTM	0.01	0.00010

Chart III - 2:


ETF
RMF
IMF
ORI DATA
Drop `Infrastracture` Industry
Drop not fully observed months
Drop outliers
Group by month, industry, type, year
Data Process
Extract Year & Month
Integrate 4 weeks into 1 monthly observation
Add `Inflow` feature
Log `AssetEnd` to make it standard distributed
Normalization
Sum the flow during the month as label
Assign label into 3 categories
Shift label 1 month late to create time series features
Features Engineering
GBDT
sale
hold
buy