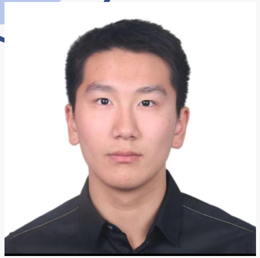
Capstone Project with Spinnaker Analytics

Team A6
Kaiyu Wang, Simeng Li,
Xiangshan Mu, Yihan Jia, Yinghao Wang

BOSTON UNIVERSITY

Our Group Members



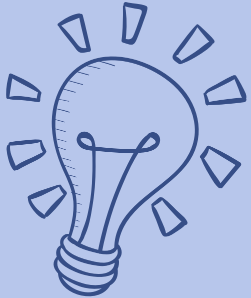
Kaiyu Wang

Simeng Li

Xiangshan Mu

Yihan Jia

Yinghao Wang



BOSTON UNIVERSITY

Table of Contents

BOSTON UNIVERSITY

01
Introduction

Overview of our analysis

BOSTON UNIVERSITY

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.

BOSTON UNIVERSITY

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
- **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

BOSTON UNIVERSITY

Project Overview

Business Objective:

- find a **tradable signal** in the dataset

Our Tradable Signal:

1. Time Series Models: Use **Flow** as output label
2. Classification Models: Use self-calculated label: **Recalculated Monthly Flow Change in Percentage**, and put it into different range buckets.

BOSTON UNIVERSITY

Business Application

- **Fund traders, mutual funds, investors and security analysts** all want a way to predict fund price fluctuations.

- Anticipating the fluctuations of fund prices (up or down) and tradable signals in advance can help fund traders make better investment decisions.

- We can also **make recommendations** based on long term, short term, different client needs and fund types.

BOSTON UNIVERSITY

02
Data Pre-processing

Basic Data Cleaning and Processing

Data Cleaning

Datasets Concatenation

Concat three datasets into one dataframe for further analysis

Add one “Type” feature to recognize category

Drop Useless Features

All the Asset and Sector values are the same, we only keep Industry

Extract Year and Month

We extract year and month from date for further EDA and classification model compiling

BOSTON UNIVERSITY

03

Exploratory Data Analysis

Help detect obvious errors,
identify outliers,
understand relationships,
find patterns within data, and
provide new insights.

BOSTON UNIVERSITY

Some Industries Have Fewer Observations

Mid Cap Growth	1737
Technology	1737
Energy	1737
Financials	1737
Health Care/Biotech	1737
Utilities	1737
Large Cap Blend	1737
Large Cap Growth	1737
Large Cap Value	1737
Mid Cap Blend	1737
Telecom	1737
Mid Cap Value	1737
Real Estate	1737
Small Cap Blend	1737
Small Cap Growth	1737
Small Cap Value	1737
Commodities/Materials	1700
Consumer Goods	1700
Industrials	1435
Infrastructure	61

Name: Industry, dtype: int64

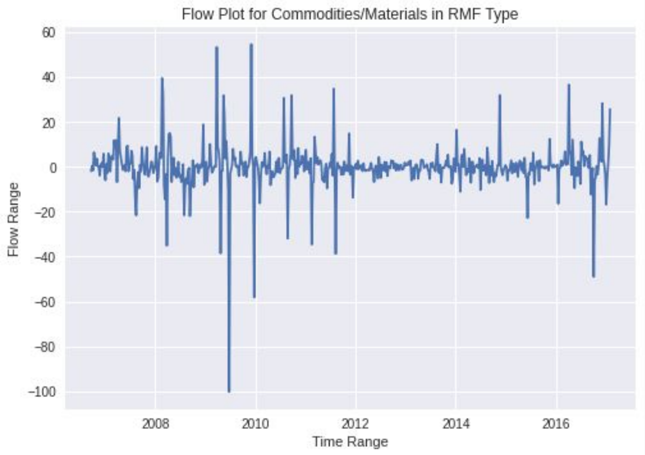
- There are **four industries have missing observations** during ten years' recording.
- We want to further analyze on these industry to understand:
  - Whether certain type of fund has fewer records than the others;
  - Where to start with missing data: at the very beginning, at the end, or discontinuity with no pattern.

BOSTON UNIVERSITY

Commodities/Materials & Consumer Goods

- Beginning of RMF Records is 37 Weeks Later Than The Other Two

```
ETF      579
IMF      579
RMF      542
Name: Type, dtype: int64
```

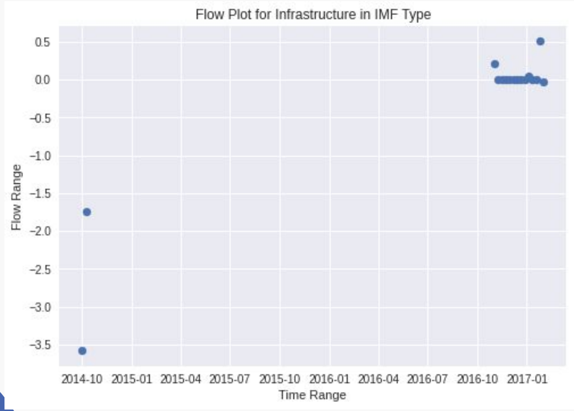


- RMF type of fund has a later beginning than the other two:
  - First record is on "2006-09-20",
  - While the other two is on "2006-01-04".
- Difference in weeks between first record of types match with number of missing value.
- There is no other discontinuity exists

BOSTON UNIVERSITY

Infrastructure Industry
- Too Few Records in Each Type

```
RMF     43
IMF     16
ETF      2
Name: Type, dtype: int64
```



- Number of observations in each type is less than 1 year.
- Infrastructure industry has a recent start for RMF and IMF:
  - RMF: the first record is on “2016-04-13”;
  - IMF: the first record is on “2014-10-01”.
- Two records of ETF all on 2014, only for 2 weeks.
- There also exists discontinuity in IMF type.

⇒ Our decision:

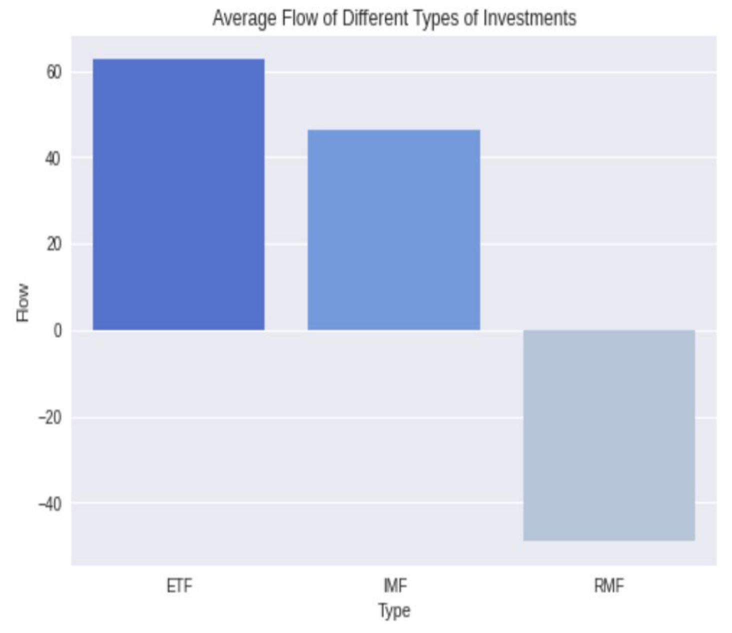
- This industry is not considered in the subsequent modeling of predicting tradable signals.

BOSTON UNIVERSITY

Average Flow among Different Types
– ETF Funds Have the Largest Average Flow

```
Max value of Flow is 23263.40489835;
Min value of Flow is -13967.24062477;
Avg value of Flow is 20.653445520794538.
```

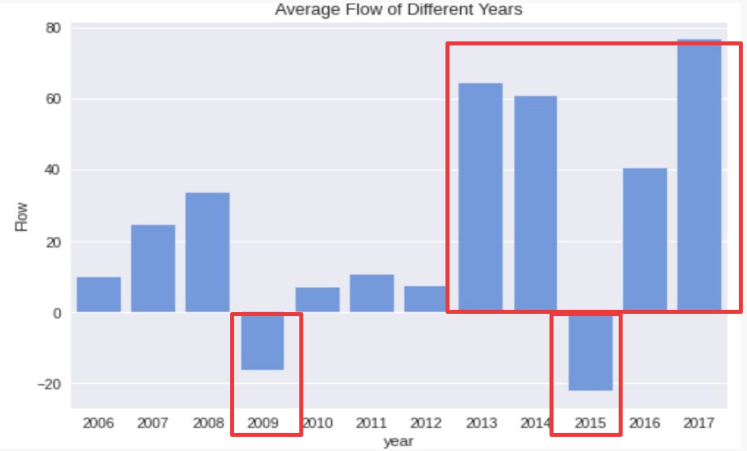
- The average flow of **ETF** funds is the largest.
- The average flow of both **ETF and IMF** funds is **greater than** the overall **average flow**.

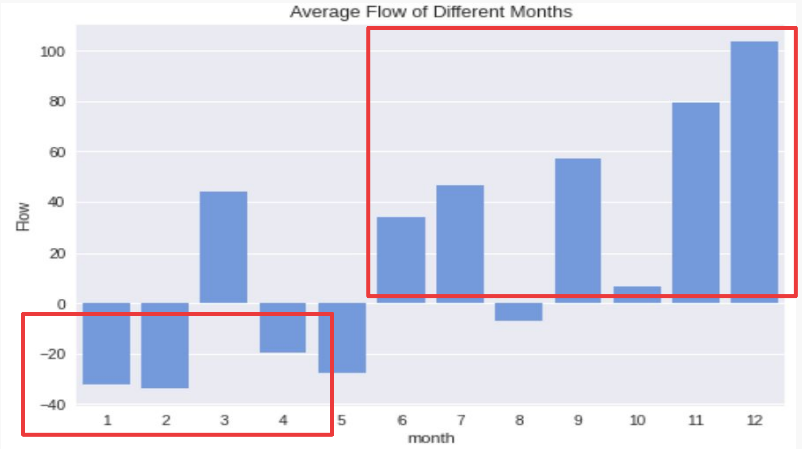


BOSTON UNIVERSITY

Average Flow among Different Years & Months

- More Inflows in the Second Half of 2013 - 2017



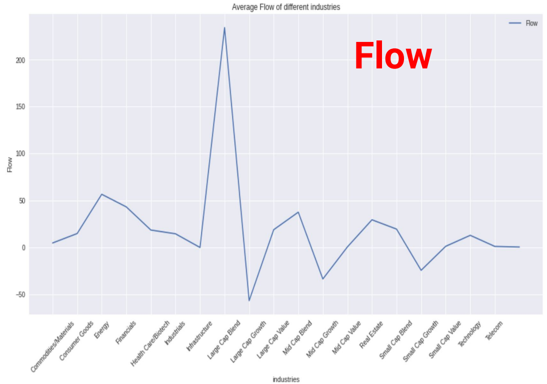


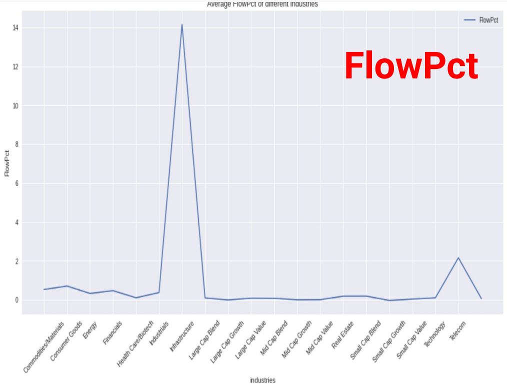
- Average flows are very high from 2013 to 2017, except for 2015.
- Average flows in 2009 and 2015 were negative and we'd guess more money was **withdrawn**.

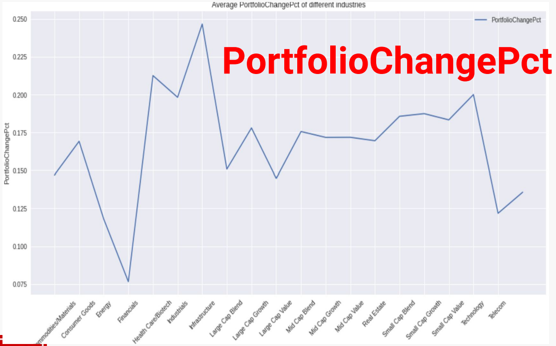
- Most of the positive average flow concentrated in the second half of the year.
- Most of the funds **flowed out** in the **first** half of the year and **flowed in** in the **second** half of the year.

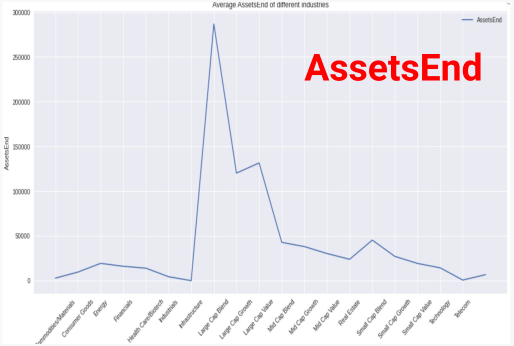
Features by Different Industries

- Flow is Less Correlated with PortfolioChangePct



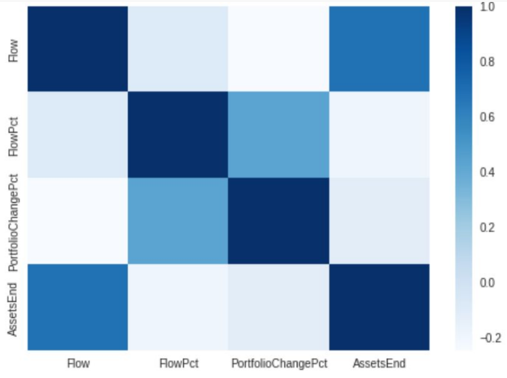






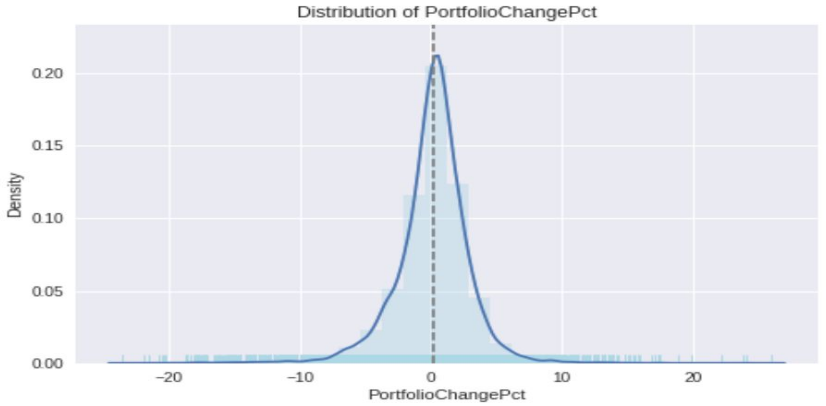
- We can see that the correlation between Flow and AssetsEnd is large.
- But the correlation between Flow and PortfolioChangePct is extremely small.

⇒ need to observe PortfolioChangePct



Average PortfolioChangePct among Different Industries

- Financial & Energy Have the Smallest Ones



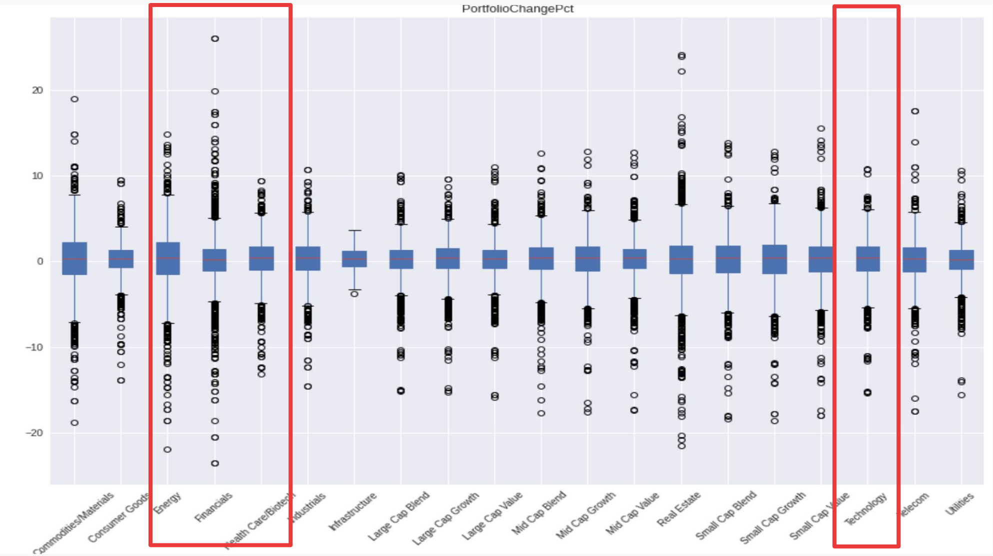
	Industry	PortfolioChangePct
3	Financials	0.076565
2	Energy	0.117991
18	Telecom	0.121771
19	Utilities	0.135717
9	Large Cap Value	0.144915
0	Commodities/Materials	0.146966
7	Large Cap Blend	0.150790
1	Consumer Goods	0.169154
13	Real Estate	0.169534
11	Mid Cap Growth	0.171738
12	Mid Cap Value	0.171795
10	Mid Cap Blend	0.175615
8	Large Cap Growth	0.177993
16	Small Cap Value	0.183333
14	Small Cap Blend	0.185696
15	Small Cap Growth	0.187447
5	Industrials	0.198256
17	Technology	0.200060
4	Health Care/Biotech	0.212619

- PortfolioChangePct is normally distributed and most concentrated in the range (-10, 10).
- Smallest average PortfolioChangePct:
  Financial & Energy → Stable → Cautious
- Largest average PortfolioChangePct:
  Health Care/Biotech & Technology → Risky → Open

BOSTON UNIVERSITY

PortfolioChangePct Distribution of Different Industries

- Decentralization Means Greater Possibility of Profit, Concentration Means Avoiding Huge Losses

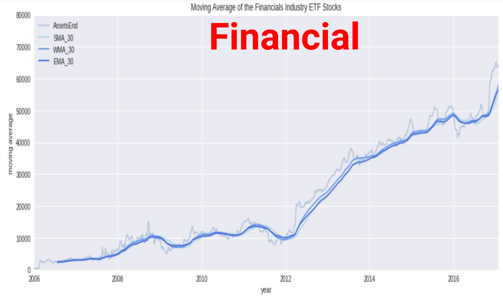


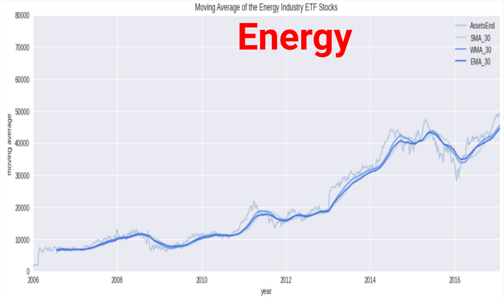
- 4 most representative industries : Financial, Energy, Health Care/Biotech & Technology
- **Financial & Energy:** the volatility is small but the distribution is the largest, which makes it easy for investors to **gain profits :)**
- **Health Care/Biotech & Technology:** The volatility is high but at the same time the distribution is small, which means that **even a drop will not bring huge losses :)**

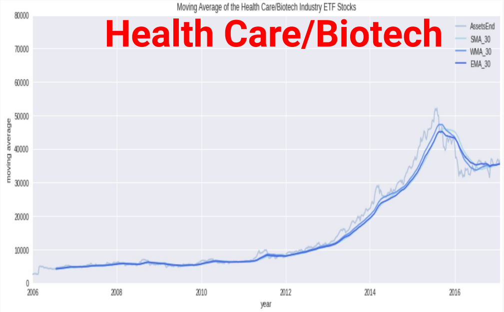
BOSTON UNIVERSITY

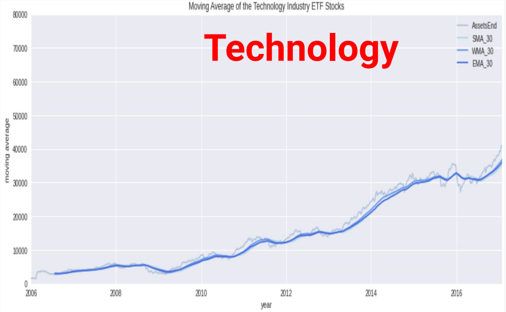
Moving Average of ETF Type

- Smoothing Fluctuations for Time Series Forecasting







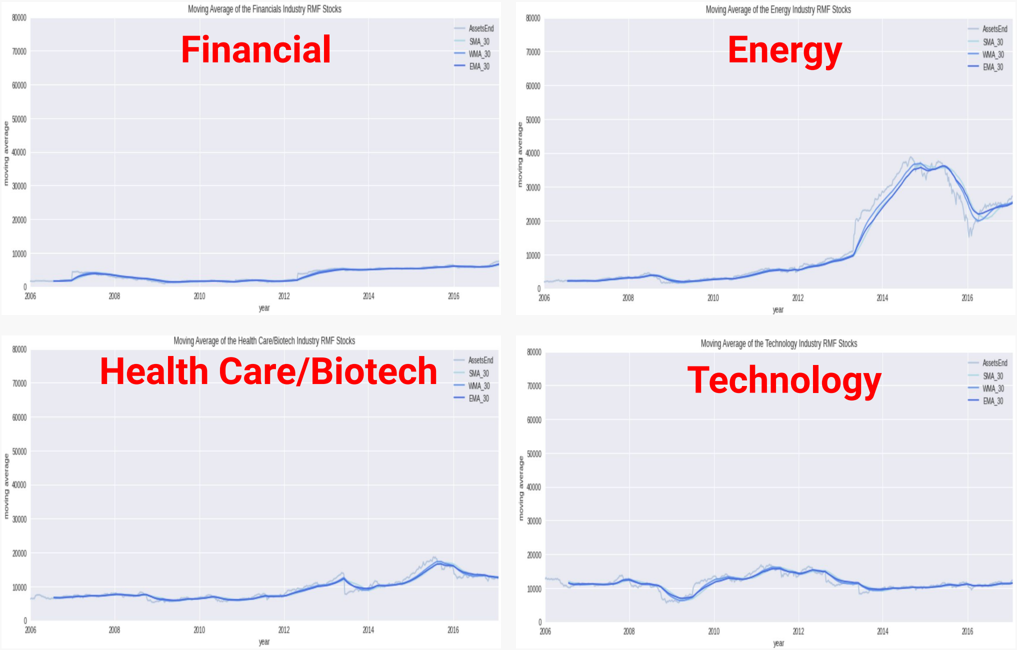


- The chart below shows **AssetsEnd** data for 4 industry ETF funds along with a **30-week** simple moving average.
- Moving averages present us with long-term volatility trends.

BOSTON UNIVERSITY

Moving Average of RMF Type

- RMF-type Funds Have the Smallest and Smoothes Values

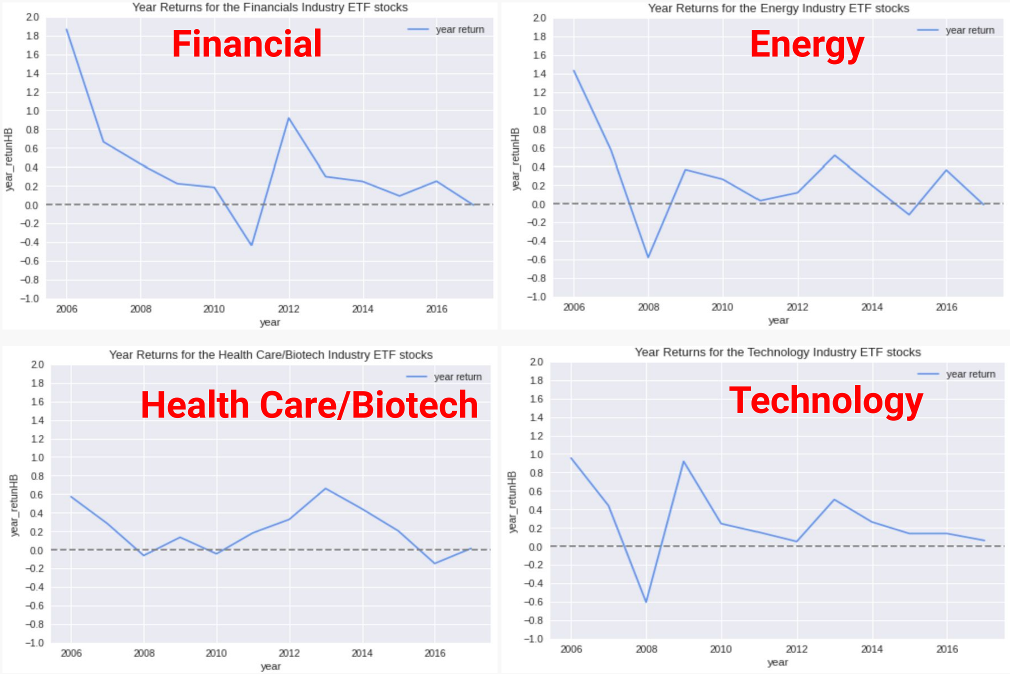


- Overall, RMF-type funds have the **smallest and smoothest** AssetsEnd values.
- And from an forecasting perspective, energy companies have the highest values of AssetsEnd.

BOSTON UNIVERSITY

Fund Year Returns on ETF Type

- Only the Health Care/Biotech Industry Shows an Increasing Trend

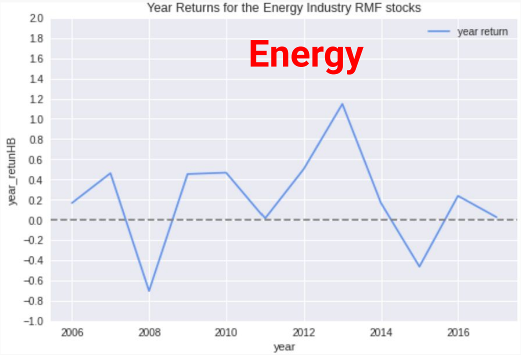


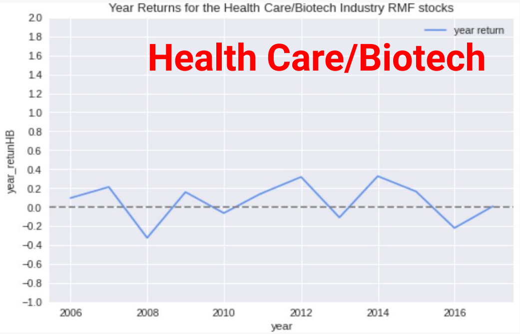
- The **Health Care/Biotech** industry has the **most stable** annual returns.
- Both 2013 and 2016 saw peak annual returns except for the Health Care/Biotech industry.
- From the recent trend observation, only the annual rate of return of the **Health Care/Biotech** industry shows an **increasing trend.**

BOSTON UNIVERSITY

Fund Year Returns on RMF Type

- Health Care/Biotech and Technology Industries Show an Upward Trend





- The annual returns of RMF funds in the **Financial and Energy** industries are more volatile and show a **downward trend** in recent years.
- The annual returns of RMF funds in the **Health Care/Biotech and Technology** industries have shown an **upward trend**.

BOSTON UNIVERSITY

Exploratory Data Analysis

- Findings & Conclusion

- **Flow** is highly correlated with AssetsEnd and Less correlated with PortfolioChangePct; More inflows in the second half of 2013 - 2017;
- **Financial & Energy:** stable, easier to profit but the annual rate of return shows a downward trend↓ [Cautious investors];
- **Health Care/Biotech & Technology:** risky, avoid huge losses, the annual rate of return↑ [Open investors];
- AssetsEnd of **IMF funds and ETF funds** move almost the same, but the IMF is more volatile; 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 **the similar trend and range** in terms of Flow for IMF and ETF type;
- Infrastructure of RMF type fund could probably be **delisted** (no need to forecast), while for the other two types we have a limited number of observations for model fitting.

BOSTON UNIVERSITY

04
Model Options

Hypothesis & solution formulation

BOSTON UNIVERSITY

Model Options

We can predict the tradable signal in **two** ways:

1. Prediction by time series regression to predict what the act.
   But it has limitations. Time series regression can only give rough predictions within a given time.
   Hypothesis test: (stationarity test)
   - h0: non-stationary, ha: stationary

   Models: MA, Exponential Smoothing, ARMA, LSTM

2. By using Multi-Class Classification Models to predict what is the range of the flow's changes.
   Models: KNN, LogisticRegression, RandomForest, Decision Tree, GBDT, AdaBoost, GaussianNB, LinearDiscriminantAnalysis, QuadraticDiscriminantAnalysis, SVM

BOSTON UNIVERSITY

Our Approach

Local Optimization
with
Time Series Analysis

Generate model for each unique combination of industry and fund type for short-term forecasting

Global Optimization
with
Multi-Class Classification

Compile a model that works for all sub-categories, can provide long-term prediction

BOSTON UNIVERSITY

05
Time Series Analysis

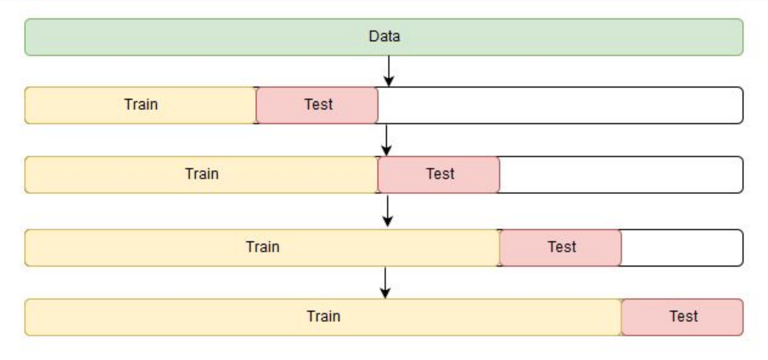
Studying the Characteristics of Flow wrt ReportDate

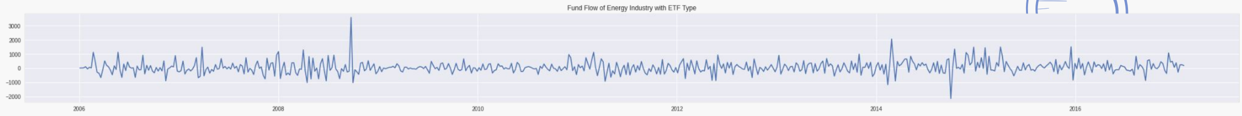
BOSTON UNIVERSITY

Time Series Analysis
- General Design for All Models

- **Model Inputs/Predictors:**
    - Past time-series data
        - ReportDate as independent variable
        - Flow as dependent variable
- **Target Variable:**
    - Flow in the future
- **Validation Method:**
    - Train test split
        - Training set: 1 ~ 550
        - Testing set: 551 ~ 579
- **Evaluation Metrics:**
    - Visually: smoothness of the flow
    - Mathematically: MSE

Forward Chaining Schematic Diagram





BOSTON UNIVERSITY

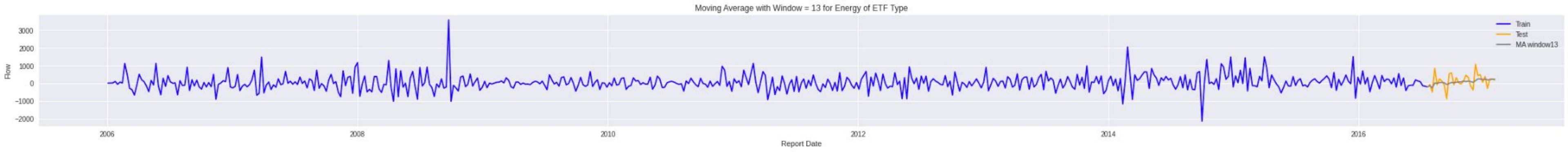
Moving Average

– MA(q) model

“Window” selection – q: $F_t = (F_{t-1} + F_{t-2} + \ldots + F_{t-q})/q$

	Type	Industry	MA window	MA_mse
0	ETF	Energy	13	1.683463e+05
1	ETF	Financials	11	5.298055e+04
2	ETF	Health Care/Biotech	14	1.037090e+04
3	ETF	Large Cap Blend	20	3.255064e+06
4	ETF	Large Cap Growth	20	1.427495e+05

- **Grid Search / Iteration** to find the best performed one:
  - Initializing Moving Average estimations with window ranged from 1 to 20,
  - Find the corresponding window that returns the minimum MSE
- Based on ACF plot:
  - Select lag index when coefficients begin to be not significant

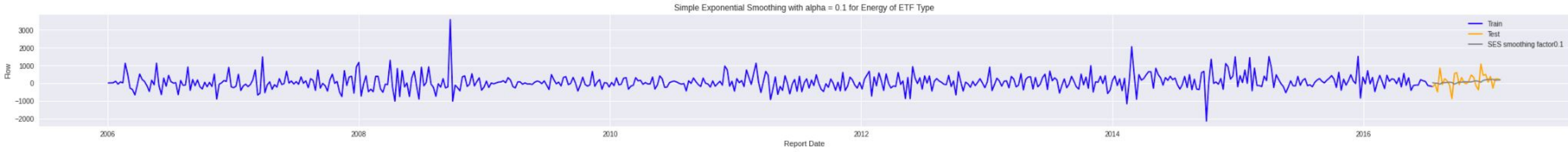


	Type	Industry	SES alpha	SES mse
0	ETF	Energy	0.1	1.810559e+05
1	ETF	Financials	0.1	5.545442e+04
2	ETF	Health Care/Biotech	0.1	1.063216e+04
3	ETF	Large Cap Blend	0.1	3.374319e+06
4	ETF	Large Cap Growth	0.1	1.492183e+05

Simple Exponential Smoothing

Smoothing level factor - α: $F_{t+1|t} = \alpha * F_t + \alpha * (1-\alpha) * F_{t-1} + \alpha * (1-\alpha)^2 * F_{t-2} + \ldots$

- Ranges from 0 ~ 1
  - 0: only the most recent values influence the forecast results
  - 1: past observations generate a larger influence on forecast results
- **Grid Search / Iteration** to find the best performed one:
  - Initializing SES estimations with alpha ranged from 0 to 1 with 0.1 step,
  - Find the corresponding alpha that returns the minimum MSE.



BOSTON UNIVERSITY

AutoRegressive Moving Average

Formula:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$

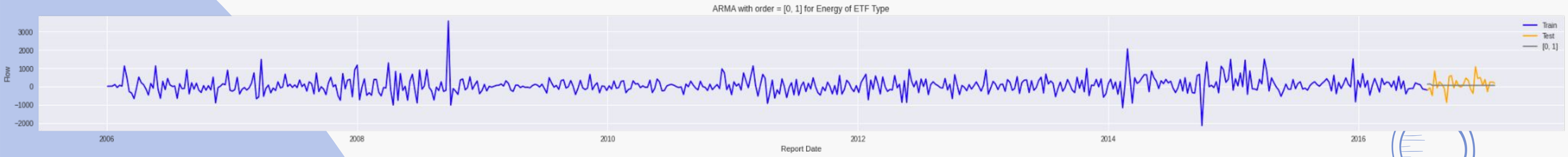
Grid Search / Iteration to find the best performed order:

- Lag Order: 0 ~ 4
- Window order: 1 ~ 4

Performance Metrics:

- Sum of AIC & BIC, the lower the better.

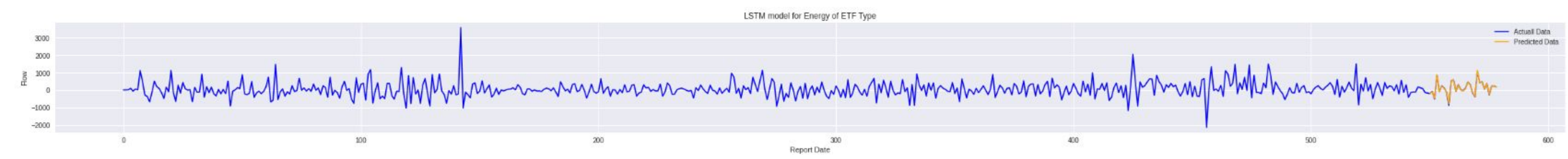
	Type	Industry	order	performance
0	ETF	Energy	(0, 1)	16550.632041
11	ETF	Financials	(3, 2)	17244.091640
19	ETF	Health Care/Biotech	(1, 2)	15711.986010
28	ETF	Large Cap Blend	(0, 1)	21365.890938
36	ETF	Large Cap Growth	(2, 1)	17625.097701



BOSTON UNIVERSITY

Long Short-Term Memory
- LSTM Model

- Can be applied to specific industry
- Follow most of jump and down
- Use the first 60 days of flowPct, AssetsEnd and PortfolioChangePct to predict flow
- StandardScaler



BOSTON UNIVERSITY

Performance Evaluation

- 2 Year Forecasting on Energy Industry of ETF Type

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

BOSTON UNIVERSITY

Time Series Forecasting
- Conclusion

Advantages:

- Simplicity, efficiency on short-term forecasting

Limitations:

- The past of the time-series data does not generate enough power to predict the future
- Three models (MA, SES, ARMA) include smoothing trends that offset the ups and downs of flow fluctuations
- Multiple additional features should be taken into account to get a more accurate forecast

BOSTON UNIVERSITY

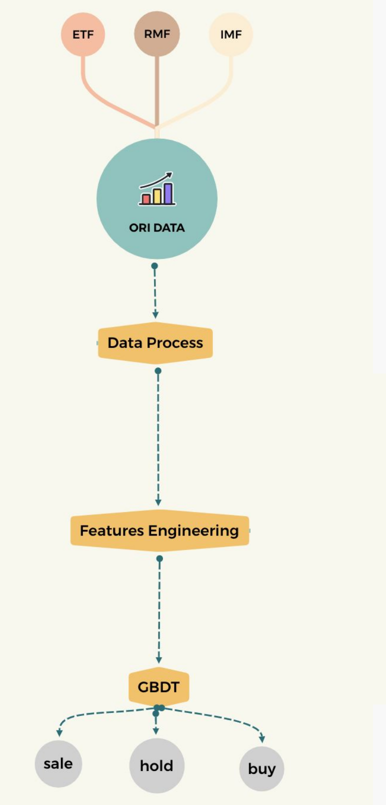
06
Classification

Multi-Class Classification Models with
Self-defined Tradable Signals

BOSTON
UNIVERSITY

WorkFlow

- The main idea of modeling at a glance



Main idea:

- To predict which interval that the next month's Flow will locate in, based on the indicators of this month.

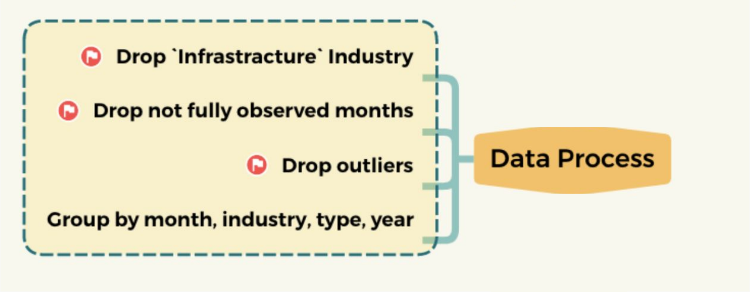
Define the tradable signal:

- Turn the monthly flow into categorical data:
    - The number of Flow located within (,-100) is 2589 (-1: sell)
    - The number of Flow located within [-100,100] is 2288 (0: hold)
    - The number of Flow located within (100,) is 2615 (1: buy)

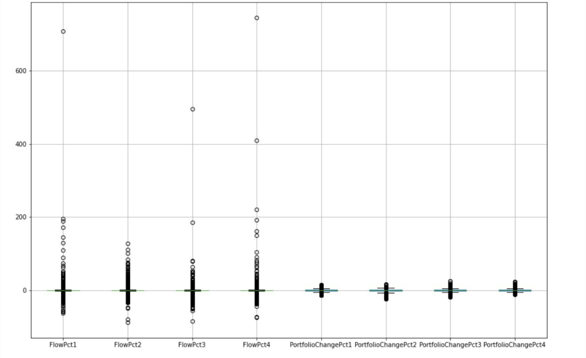
BOSTON UNIVERSITY

Data Processing

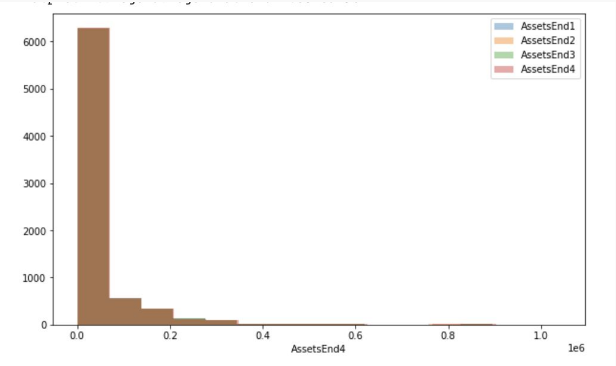
- After the mid-term, we did extra in. . .



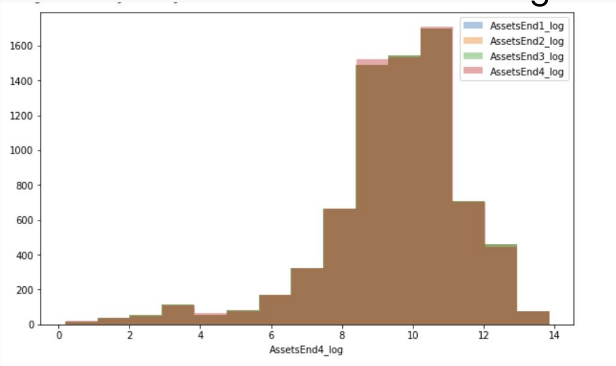
Check outliers of PlowPct & PorfolioChangePct



Distribution of AssetEnd



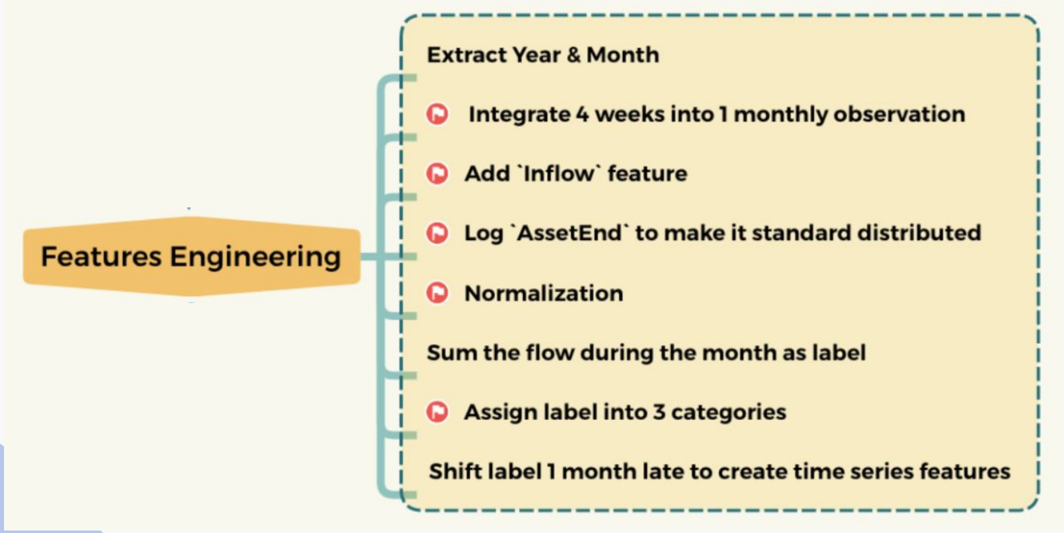
Distribution of AssetEnd after log



BOSTON UNIVERSITY

Features Engineering

- After the mid-term, we did extra in. . .



Final Size of the data:

- 7493 rows × 11 columns

Sampling:

- Split train and test into 80% and 20%

BOSTON UNIVERSITY

Features Engineering

	ReportDate	AssetClass	Flow	FlowPct	AssetsEnd	PortfolioChangePct	Type
0	2/1/2017 12:00:00 AM	Commodities/Materials–North America–USA–North ...	378.578706	4.5064	8679.056347	–1.1938	ETF
1	2/1/2017 12:00:00 AM	Consumer Goods–North America–USA–North America...	332.526792	1.1479	28973.613065	–1.1231	ETF
2	2/1/2017 12:00:00 AM	Energy–North America–USA–North America–Equity	186.031374	0.3782	48446.700077	–1.8855	ETF

32688 rows × 7 columns:

- Contact 3 dataset
- Process AssetClass

	ReportDate	Flow	FlowPct	AssetsEnd	PortfolioChangePct	Type	Industry
0	2017–02–01	378.578706	4.5064	8679.056347	–1.1938	ETF	Commodities/Materials
1	2017–02–01	332.526792	1.1479	28973.613065	–1.1231	ETF	Consumer Goods
2	2017–02–01	186.031374	0.3782	48446.700077	–1.8855	ETF	Energy

32688 rows × 7 columns: still

- Drop Infrastructure industry
- Drop not fully observed months
- Drop outliers
- Extract year and month

	ReportDate	Flow	FlowPct	AssetsEnd	PortfolioChangePct	Type	Industry	year	month
10667	2006–01–04	–9.076384	–0.2789	3293.597159	1.5210	RMF	Utilities	2006	1
10989	2006–01–04	5.144223	2.4847	213.485252	0.6311	ETF	Industrials	2006	1
10988	2006–01–04	45.038766	1.7847	2601.925992	1.3221	ETF	Health Care/Biotech	2006	1

32564 rows × 9 columns:

- Integrate into monthly observed

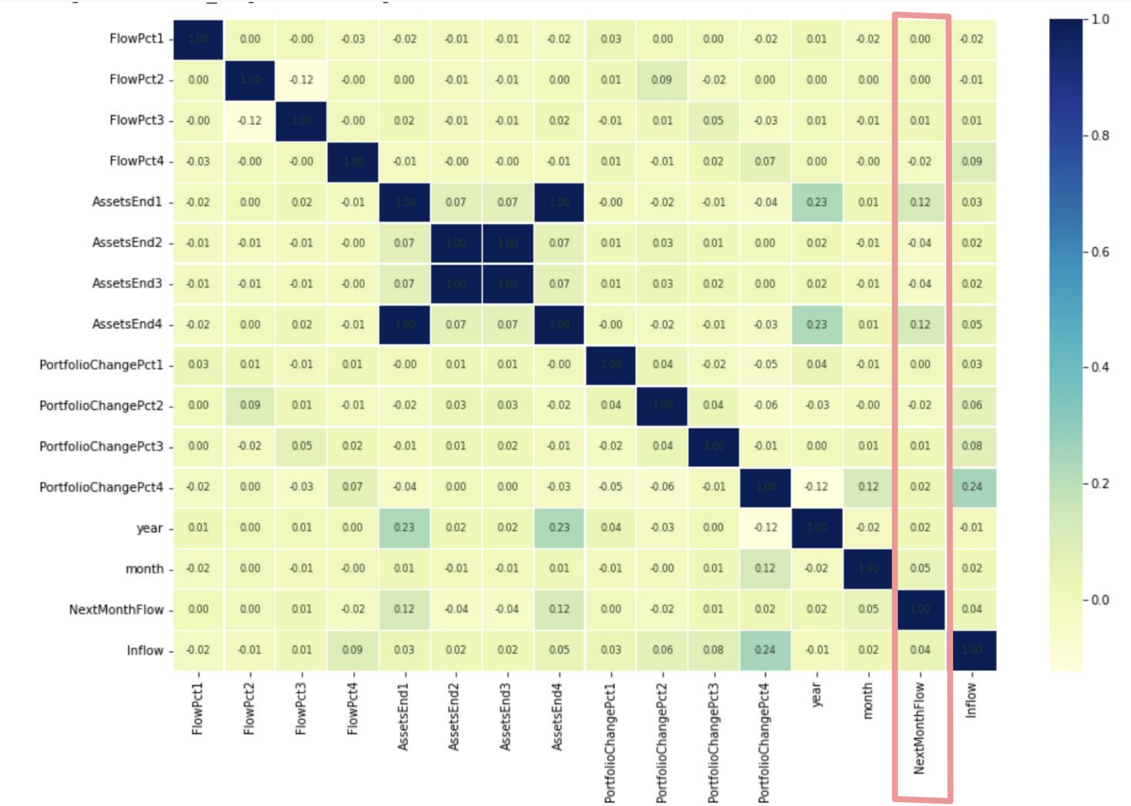
7493 rows × 11 columns

Industry	Type	year	month	FlowPct1	FlowPct2	FlowPct3	FlowPct4	AssetsEnd1	AssetsEnd2	AssetsEnd3	AssetsEnd4	PortfolioChangePct1	PortfolioChangePct2	PortfolioChangePct3	PortfolioChangePct4	Type	Industry	year	month	NextMonthFlow
Commodities/Materials	ETF	2006	1	6.8877	3.1609	–0.6022	–9.5029	365.747363	376.895214	367.731276	342.463192	1.9393	–0.1130	–1.8291	2.6315	ETF	Commodities/Materials	2006	1	39.448926
			2	0.0000	11.1578	0.0000	11.1578	353.551743	380.566108	382.711351	389.790653	3.2378	–3.5170	0.5636	1.8497	ETF	Commodities/Materials	2006	2	165.453872
			3	0.6584	–6.9428	–4.0455	4.0095	1226.034803	1113.668457	1119.426080	1426.387768	–0.7101	–2.2222	4.5625	1.9107	ETF	Commodities/Materials	2006	3	–134.278695
			4	–6.7927	–2.7356	2.0618	–1.9407	1360.070894	1307.920822	1398.830639	1360.132199	2.1434	–1.0987	4.8888	–0.8257	ETF	Commodities/Materials	2006	4	4.147121
			5	–5.8893	3.0092	–0.2588	2.9431	1286.578357	1361.731888	1254.982939	1293.819947	0.4815	2.8320	–7.5803	1.3907	ETF	Commodities/Materials	2006	5	299.481155

BOSTON UNIVERSITY

Correlation Heatmap

- **Focus on the correlation with the trading signal**



- Add more predictors
- For example, the 2nd and 3rd week's FlowPct, AssetsEnd and PortfolioChangePct in the month
- The data are more powerful in predicting the tradable signal

BOSTON UNIVERSITY

Model Evaluation

- Comparing with the mid-term

Best Model in Midterm

```
Start Training: GBDT

accuracy is 0.538970
```

	precision	recall	f1-score	support
1	0.62	0.60	0.61	398
2	0.57	0.43	0.49	370
3	0.44	0.43	0.43	333
4	0.52	0.67	0.59	413
accuracy			0.54	1514
macro avg	0.54	0.53	0.53	1514
weighted avg	0.54	0.54	0.54	1514

Best Model in Final

```
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]]
```

- Fit the data into 10 classifiers from sklearn module
- BEST classifier: GBDT
- Comparing with midterm, the Final model has increased 25% in the overall accuracy
- Precision on “Sell” is increased 3.2%
- Precision on “Hold” is increased 48%
- Precision on “Buy” is increased 23%

BOSTON UNIVERSITY

Multi-class Classification

- Conclusion

- One advantage of classification is we can give **a global optimization using only one model**.

- **Limitations:**
  - We did not consider the currency value has changed during the past 10 years. Change the trading signal as ratios might be a solution.
  - More features are needed to show the performance of the stock, from which the model can capture more information.

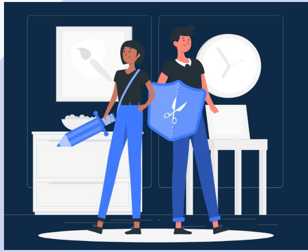
BOSTON UNIVERSITY

07
Use-Case Scenario

How our models help

BOSTON
UNIVERSITY

Business Application
- Time Series Analysis

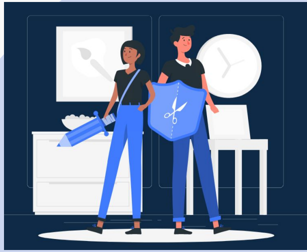


With our Time Series Analysis:

- We are able to find funds which are more stable in flow amount, less volatile to market, and less sensitive to time; these funds are concluded to be the **safety choice to invest**.
- Asset managers can use our findings to **classify funds** (stable/volatile)
- Based on the MSE output, LSTM gives the highest accuracy on **forecasting near future Flow value**.
- With our forecast on future funds' value, we can give clients' suggestions on **whether to invest** when flow decreases, **whether to sell** on-hand type when flow increases.

BOSTON UNIVERSITY

Business Application
- Classification Analysis

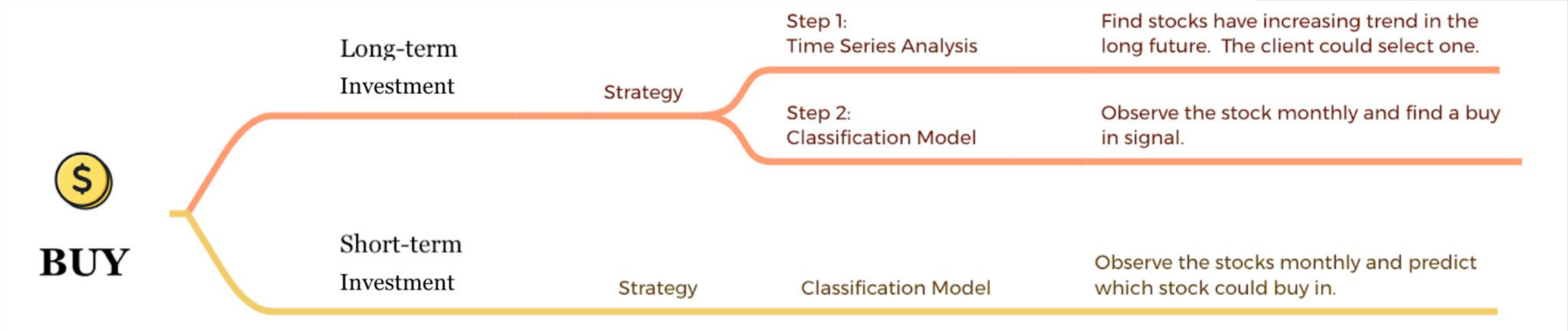


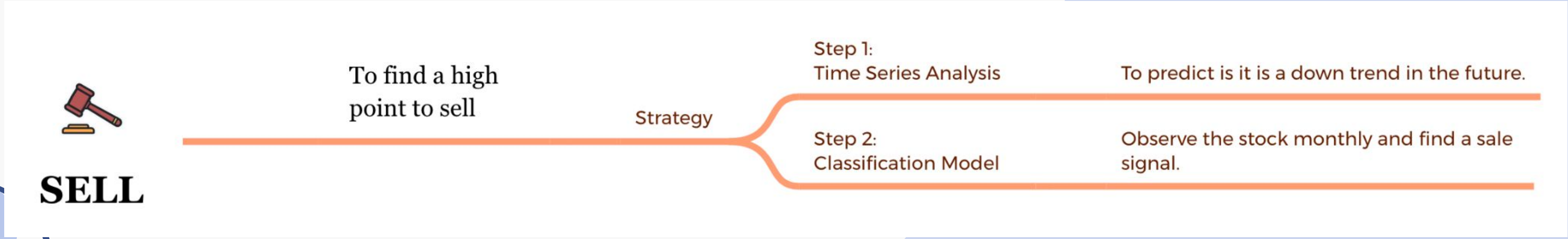
With our Classification Analysis:

- Help clients **predict the specific Flow range for the future month**.
- Help clients **choose a fund** in the ideal Flow range.
- Forecast the trend of **fund market fluctuation**.

Business Application

- Buy & Sell





BOSTON UNIVERSITY

THANKS

Any questions?

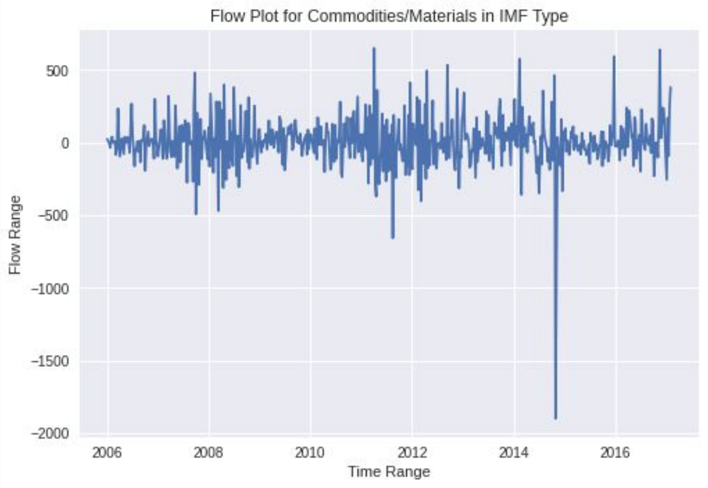
BOSTON UNIVERSITY

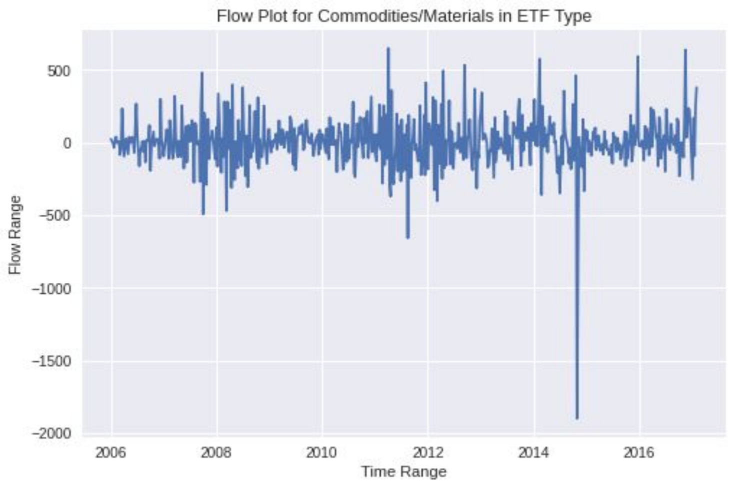
Appendix

Additional Materials/Findings of our project

Commodities/Materials Industry

- IMF and ETF Share Similar Pattern in terms of Flow





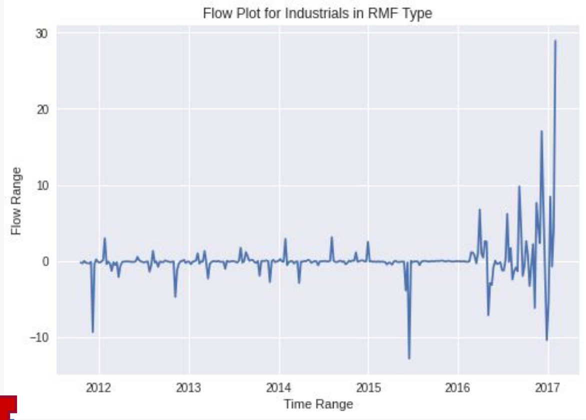
- We can observe a very similar pattern of flow rises and drops;
- Not only did the changes happen at the same time, but the magnitudes of increases or decreases were roughly the same.

BOSTON UNIVERSITY

Industrials Industry

- Have only 5 years of RMF Records

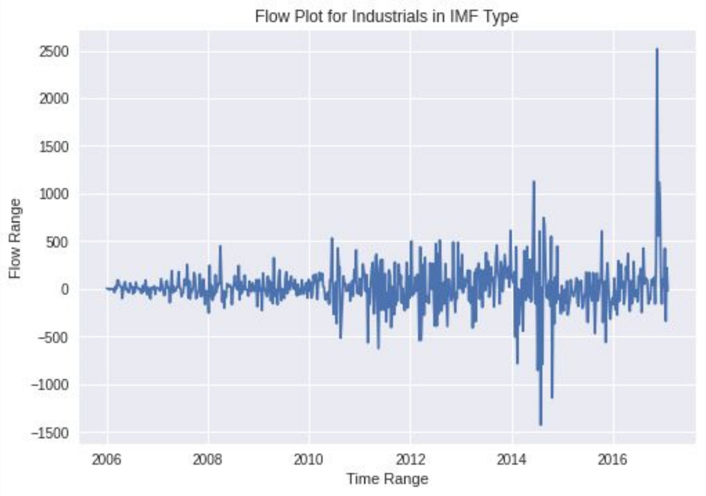
```
ETF     579
IMF     579
RMF     277
Name: Type, dtype: int64
```

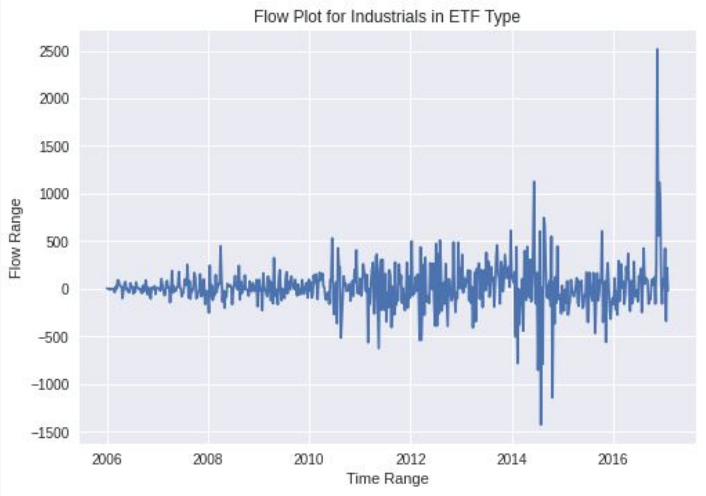


- RMF type of fund has a later beginning than the other two:
    - First record is on “2011-10-19”,
    - While the other two is on “2006-01-04”.
    - Meaning we have only about 5 years of RMF records, which is half less than ETF or IMF.
- Difference in weeks between first record of types match with number of missing value.
- There is no other discontinuity exists

Industrials Industry

- IMF and ETF Share Similar Pattern in terms of Flow



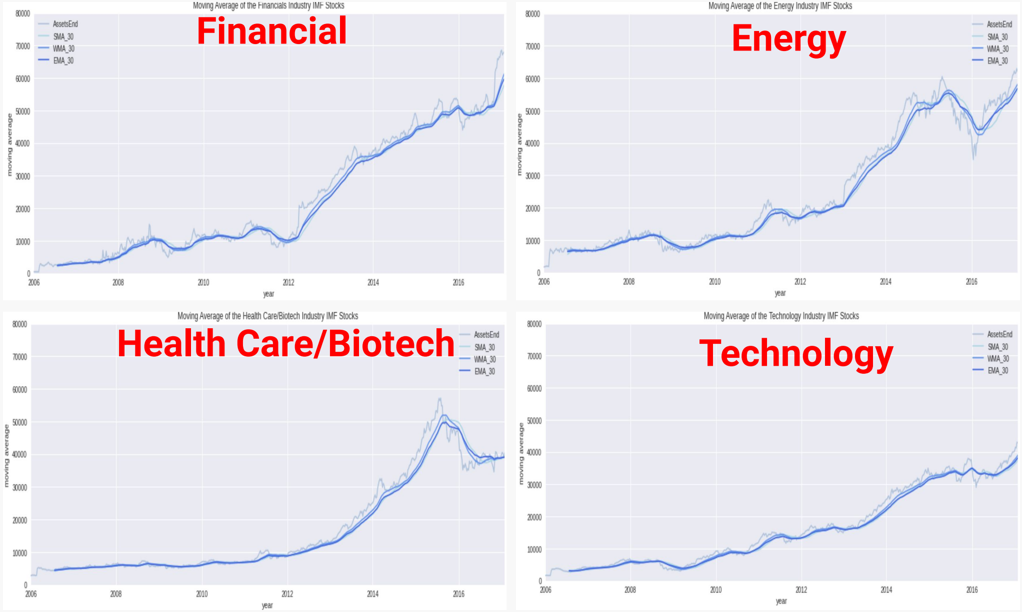


- We can observe a very similar pattern of flow rises and drops;
- Not only did the changes happen at the same time, but the magnitudes of increases or decreases were roughly the same.

BOSTON UNIVERSITY

Moving Average of IMF Type

- Volatility is a Little Bit Larger than that of ETF

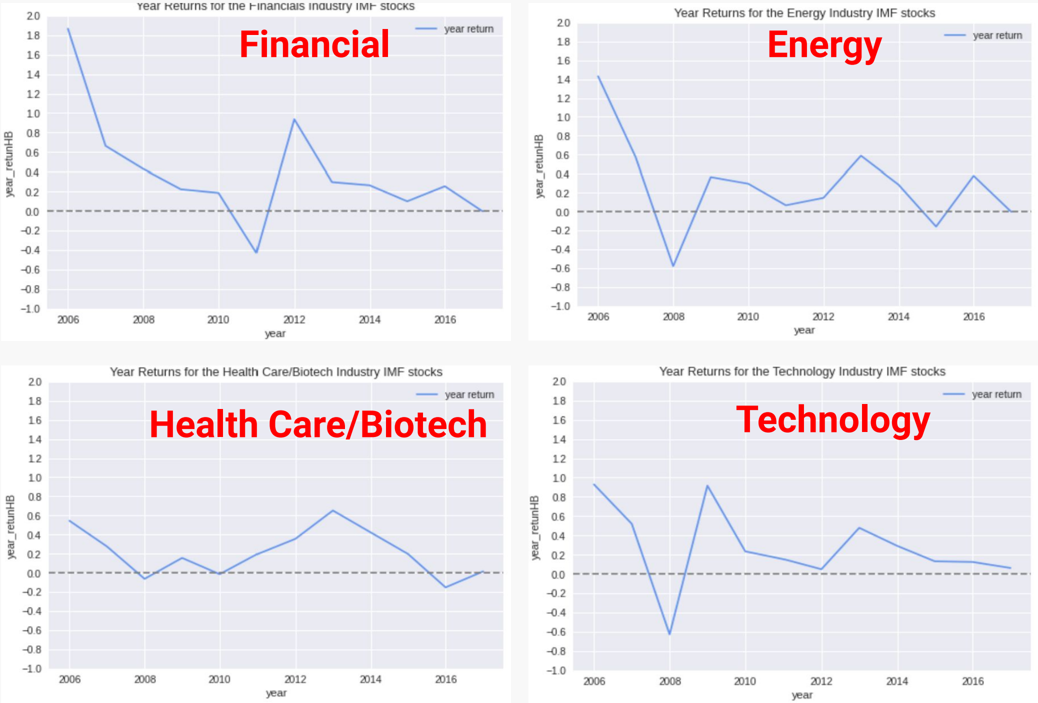


- In general, the trend is **almost the same** as that of ETF funds AssetsEnd.
- But all the volatility is a little bit larger than that of ETF.

BOSTON UNIVERSITY

Fund Year Returns on IMF Type

- No Big Difference Compare with the ETF Type

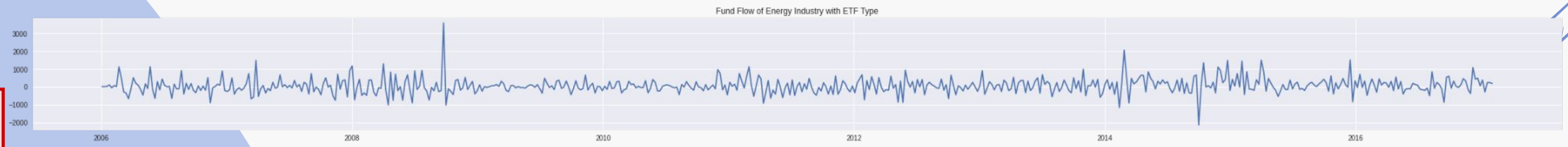


- There is **no big difference** between IMF fund annual returns and ETF returns.

BOSTON UNIVERSITY

Time Series Analysis
- Basic Properties

- **Trend:**
  - Refers to a long-term increase or decrease in the time-series data
  - Can be a roughly slope going through the data
- **Seasonality:**
  - Refers to periodic fluctuations.
  - If there exists repeated pattern, we can intuitively observe the length of the interval from the line chart of flow by time.
- **Cyclicity:**
  - Refers to (regular or periodic) fluctuations around the trend, excluding the irregular component, revealing a succession of phases of expansion and contraction.

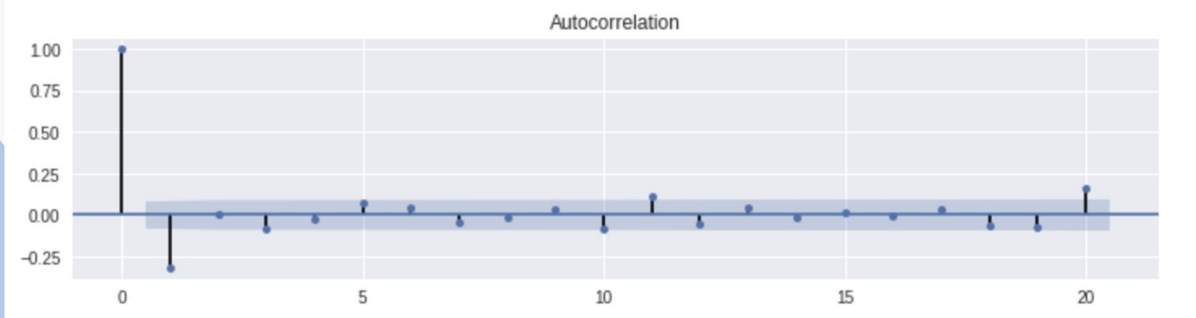


BOSTON UNIVERSITY

Time Series Analysis

- Relationship between current value and past

- **Autocorrelation:**
  - Refers to the similarity between observations as a function of the time lag between them.
  - We can also derive the inference about seasonality from autocorrelation plots.
  - 1 spike outside the blue region, meaning that there is correlation at lag 1.
  - Difficult to identify a seasonality, because increases and decreases are not showing a clear pattern of same interval.



BOSTON UNIVERSITY

Time Series Analysis

- The Golden Rule

- **Stationarity:**
  - Check whether statistical properties such as mean, variance, and covariance do not vary with time.
  - Why Important: stationarity is the pre-requirement of a time-series data continuing to “inertia” along the existing form in a further period, thus, it is the precondition for predictable.
  - Each subset is proved to be **stationary** through the the ad-fuller test, with this as a premise, we can move on with MA, SES and ARMA estimations.

```
Stationary check for type ETF in Energy industry:
ADF Test Statistic : -21.07813363393672
p-value : 0.0
#Lags Used : 1
Number of Observations : 577
Strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data is stationary.
```

BOSTON UNIVERSITY

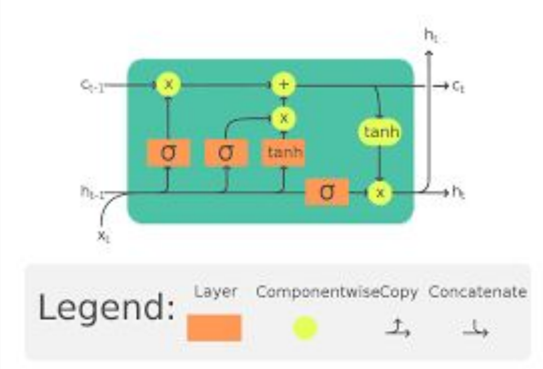
Time Series Analysis
- Models From Simple to Complex

1. **Moving Average:**
   - States that the next observation is the mean of all past observations.
   - A smoothed flow line with highlighted different trends.
2. **Exponential Smoothing:**
   - Different weights are assigned to different observations: the more recent, the more important,
3. **ARMA:**
   - AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.
   - MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

BOSTON UNIVERSITY

Time Series Analysis
- Models From Simple to Complex Cont.

4. **LSTM:**
   - Long Short Term Memory networks are a special kind of RNN, capable of learning long-term dependencies.
   - Frequently in used for Time-Series Prediction
   - **LSTMs can be used where ever there is time dependent information**.
     (To be precise wherever there is sequence data)



Assess MA* vs SES*

	Type	Industry	MA window	MA_mse	SES alpha	SES_mse
0	ETF	Energy	13	1.683463e+05	0.1	1.810559e+05
1	ETF	Financials	11	5.298055e+04	0.1	5.545442e+04
2	ETF	Health Care/Biotech	14	1.037090e+04	0.1	1.063216e+04
3	ETF	Large Cap Blend	20	3.255064e+06	0.1	3.374319e+06
4	ETF	Large Cap Growth	20	1.427495e+05	0.1	1.492183e+05

MA Returns smaller MSE when the window value is larger:

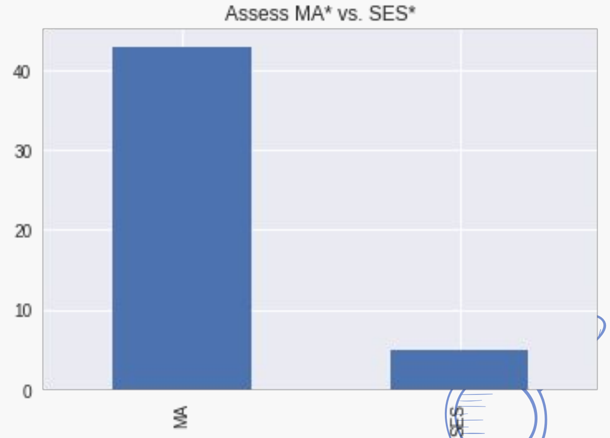
- **Underlying pattern of flow is stable**

SES returns smaller MSE when smoothing trend is smaller:

- **Flow forecast is less sensitive**

For most local optimization approaches:

- **MA performed better** than SES



BOSTON UNIVERSITY