Predicting Trading Signals

May 6th, 2022

Capstone Team B1:
Carmen Cheung, Song Lin, Haiyuan Zhang, Linh To, Zichen Wang

Table of Contents

1. Abstract

The purpose of our project is to identify the tradable signal based on the weekly funds flow data. We are trying to fit and compare different models to capture the market trend in future and define a proper threshold for signals.
For some of the models, we preprocessed the data to remove the effect of market risk exposure and some noises in our data.
Based on our definition of tradable signal, we finally settled on two models which performed the best of all the models we tried, which is SARIMAX and DES. More detailed information will be provided in the modeling section.

2. Introduction

The purpose of this project and analysis is to analyze fund flow data from three different types of investments and identify tradeable signals that will aid in investment strategies. A multitude of modeling tools are used to build selected suitable models to tackle this problem and compare the results to eventually recommend the best model and its uses.

3. Dataset Overview

3.1 Overview

The dataset contains weekly data of 3 types of investments: institutional ETF, institutional mutual fund, and retail mutual fund, from 2006 to 2017 with the following variables.

3.2 Variable Description

- ReportDate: Weekly data from 2006 to 2017
- AssetClass: Name of the portfolio
- Flow: How much money ($M) is coming in or cashing out
- FlowPct: Flow/Assets beginning of the week *100%
- AssetsEnd: Assets at the end of the week ($M)
- PortfolioChangePct: Percent change in overall portfolio during the week

4. Data Cleaning

4.1 Missing Values

While some asset classes had less observations than others (due to not recording data for that asset class earlier on), almost all industries have no missing values within their continuous date of recorded data. The only asset class with missing data was infrastructure. Due to the large number of missing observations and the fact that the datasets still have 19 other industries to analyze, we decided to drop this asset class from our analysis.

5. Exploratory Data Analysis

5.1 Average PortfolioChangePct Trend

In a plot of average portfoliochangepct over time, all industries have a similar trend shape with a noticeable observed dive around October 2008 and an upward trend around April 2016. This suggests that there may be factors that impact the entire market. Systematic risk may be a significant factor in

how funds move and thus if we eliminate the influence of the general economy, we may be able to extract the residual impact attributed to that industry.

5.2 Flow Trend

The flow variable for ETF and MF have a similar pattern over the past 10 years in most asset classes. This suggests that these two investment portfolios have similar behaviors. The retail mutual fund flow showed a similar trend initially, but from 2012 on, the flow change is relatively shrinked compared to institutional flow trends. In late 2014 to early 2015, there are observed high flow fluctuations amongst retail investors.

5.3 AssetEnd Trend

The assetend variable trend for institutional investors in both ETF and mutual funds look very similar and show a steady increase trend in the past 10 years. However, for the individual investors, the assetend change experiences more fluctuations across the asset classes.

6. Hypotheses

From EDA, we formulated four hypotheses to guide our analytical approach:

1. Using a moving average to smooth the trend can help indicate future patterns.
2. Removing systematic market factors using the S&P 500 as a proxy could indicate which industries are performing better or worse relative to the market
3. Time-series models are expected to perform the best since the order of the data matters in predicting future values
4. Each industry may have unique inherent features that can be explained better with different models according to each model's strength.

7. Label Generation

We do not have existing labels to mark tradable signals, so for the following models, we need to design labels and corresponding thresholds ourselves.

First, we want to eliminate as much noise as possible from the data. One of the most simple methods is to use a moving average over a longer period.

In addition, according to the CAPM model, stock market returns can be decomposed into returns affected by systematic risk and returns affected by idiosyncratic risk. Feeding the data into the predictive model after preprocessing the CAPM model may improve the performance of our model.

7.1 Rolling Average of Asset-End

We compare the average value of assetsend of future 25 weeks with the target date's asset end, so as to eliminate the noise caused by the short-term fluctuation in the future as much as possible.

In this example, we choose ±5% as the threshold. When the average asset end value for the next 25 weeks is 5% or more higher than the current asset end value, the current date is labeled as 1, which means there is a future growth trend. When the ending asset average for the next 25 weeks is 5% or more lower than the current asset average, the current date is labeled as -1, which means there is a future downtrend. If within plus or minus 5%, we consider the market to be flat and marked as 0.

The distribution of labels produced is: 52% for 1, 33% for 0, and 15% for -1. It shows that the market has been in a long-term upward trend for more time, which is also in line with the general trend of the stock market.

7.2 Removing Systemic Factors

In this method, we use the S&P 500 to measure the level of market returns, and use the CAPM model to regress the data for each asset class separately. The difference between the regression value and the actual value reflects the stock returns that are affected by idiosyncratic risk.
After calculating the residual term, we can either use the moving average of a future period to measure the future stock market trend of each industry, or use methods such as t-test to test its statistical significance and use it as one of the labeling criteria.
At the 95% confidence level, 12% of the proportions have future returns that are statistically significantly different from 0, independent of systematic risk.

8. Classification Modeling

8.1 Logistic Regression

We first tried logistic regression since the goal of the project was to predict trade signals which is a classification problem. However, logistic regression was capable of handling time-series data but we still wanted to include it as a baseline for our project.

Training and Predicting:

Train-test ratio: 80-20
Target variable: "Label" column created by rolling average of Asset End

- We took the rolling average of column "AssetEnd" for every 25 weeks and created a new column called "Rollingaverage"
- The reason we decided on 25 weeks is we want to predict half a year value
- We calculated trading signal with (rollingaverage-AssetEnd) / AssetEnd , if it's > 5% then it's a buy signal, if it's between -5% and 5% then it's a hold signal, it it's < -5% then it's a sell signal

Predictors: Flow, FlowPct, PortfolioChangePct

Fit the model and predict: We ran the model through 3 datasets to get the trading signal

Evaluation:
Model result:

Report for etf				
	precision	recall	f1-score	support
-1	0.2	0.29	0.24	349
0	0.49	0.33	0.4	729
1	0.54	0.56	0.55	1123
accuracy			0.44	2201
macro avg	0.41	0.4	0.39	2201
weighted avg	0.47	0.44	0.45	2201

Report for institutional				
	precision	recall	f1-score	support
-1	0.17	0.23	0.2	305
0	0.49	0.57	0.53	798
1	0.56	0.45	0.5	1098
accuracy			0.46	2201
macro avg	0.41	0.41	0.41	2201
weighted avg	0.48	0.46	0.47	2201

Report for retail				
	precision	recall	f1-score	support
-1	0.31	0.14	0.19	419
0	0.55	0.58	0.56	949
1	0.41	0.52	0.46	757
accuracy			0.47	2125
macro avg	0.42	0.41	0.41	2125
weighted avg	0.46	0.47	0.45	2125

9. Time-Series Predictive Modeling

9.1 VAR Model - Cross-Market without Moving Average

We also tried the Vector AutoRegression (VAR) model, with the underlying logic that each variable is a linear function of past lags of itself and past lags of the other variables. The key point of our hypothesis is to reveal that the AssetEnd could be predicted by the other variables in different markets. We constructed the VAR model 60 times on 20 asset classes of 3 datasets (ETF, Institutional and Retail).

Function:

- *%AssetEnd(t) = a + b1* %Flow(t-1) + b2* %FlowPct(t-1) + b3* %PortfolioChangePct(t-1)*

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.

Training and Predicting:

Train-Test Ratio: Test on last 10 weeks
Target variable: AssetEnd
Predictors: Flow, FlowPct, PortfolioChangePct

Fit the model and predict

We divided the 3 datasets 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. We divided the 3 datasets with 20 asset classes in each into 60 small datasets, and run the loop through the predictors to predict the last 10 weeks of AssetEnd

Evaluation:

The model result:

MSE with VAR

	Etf	Institutional	Retail
model_Utilities	169229	220717	230438
model_Commodities/Materials	520807	532447	1603336
model_Consumer Goods	137509	137509	156514
model_Energy	7137668	12140902	143104
model_Financials	45174061	49626576	10004292
model_Health Care/Biotech	657868	1289787	522038
model_Industrials	2464983	2464983	44807894
model_Large Cap Blend	693291565	1201883001	7417725
model_Large Cap Growth	36367426	11874972	254
model_Telecom	69397	69397	733402
model_Mid Cap Blend	22428921	46015415	1236547
model_Technology	4179707	4336722	5741649
model_Small Cap Value	5296657	13461327	222554
model_Small Cap Growth	968943	1517035	561745
model_Large Cap Value	21254147	80403700	63447
model_Real Estate	2754912	2507222	110281
model_Mid Cap Value	1566044	7344058	1238
model_Mid Cap Growth	122381	27906	2289
model_Small Cap Blend	75708490	133372346	771

VAR performed the best with Large Cap Growth and Small Cap Blend for Retail dataset with an MSE of 254 and 771, but not with other asset classes

9.2 SARIMAX Model - with Seasonal Component

Seasonal Autoregressive Integrated Moving Average, SARIMA , is an extension of ARIMA that **explicitly supports univariate time series data with a seasonal component.**

It adds three **new hyperparameters** to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

Training and Predicting:

Train-test ratio: 80-20
Target variable: AssetEnd
Predictors: Flow, FlowPct, PortfolioChangePct, M16, M50

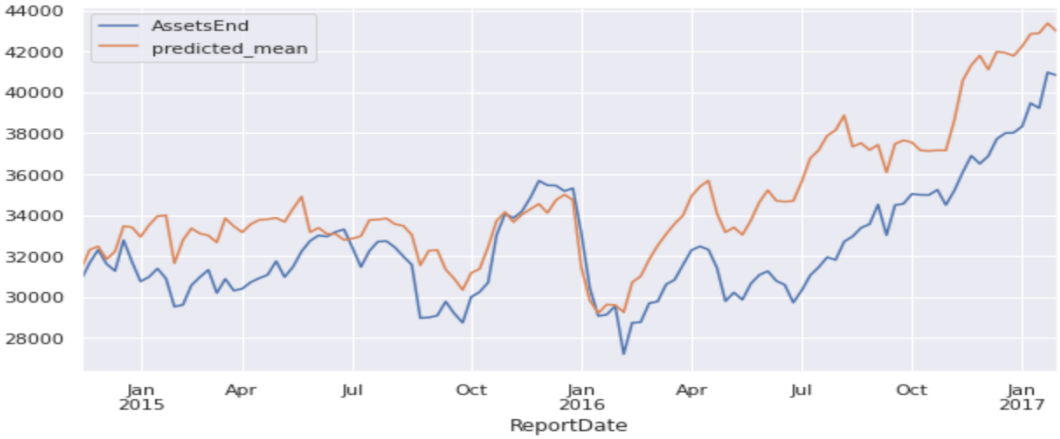
Besides predictors in the VAR model, we performed feature engineering to generate additional features for this model. These features are derived from the given dataset itself. These features also highlight some common trends that helps the model to predict better:

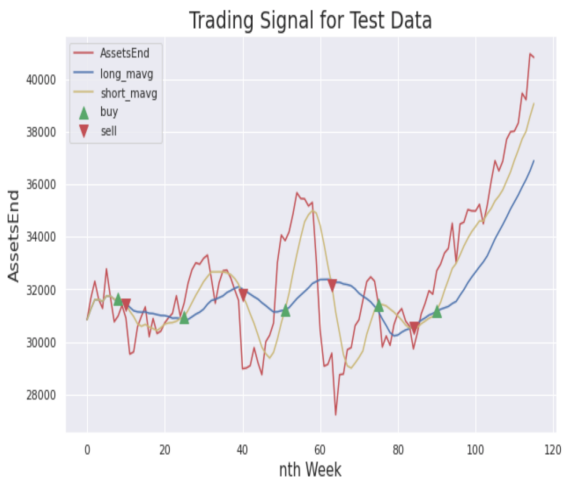
- **M16** : Moving Average 16 M16 uses a window of 16 weeks and calculates the values based on the past 16 moving average of the AssetsEnd.
- **M50** : Moving Average 50 M50 uses a window of 50 weeks and calculates the values based on the past 50 moving average of the AssetsEnd.

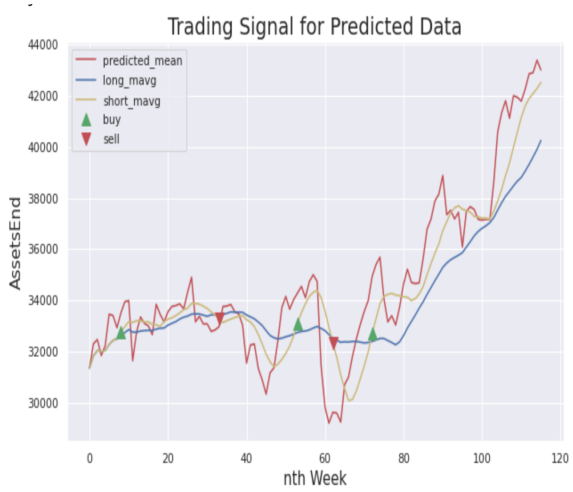
After tuning the model with appropriate parameters, we defined trading signal classification based on the model training and testing result. We set a short moving average feature as 8 weeks and long moving average feature as 20 weeks in both test and predicted dataset. A crossover occurs when a short moving average crosses a long moving average. In stock trading, this meeting point can be used as a potential indicator to buy or sell an asset.

- When the short term moving average crosses above the long term moving average, this indicates a buy signal.
- When the short term moving average crosses below the long term moving average, it may be a good moment to sell.

Furthermore, we built an easy function for users to test prediction on any asset class in each dataset. The following example is Technology asset class prediction summary on ETF dataset:







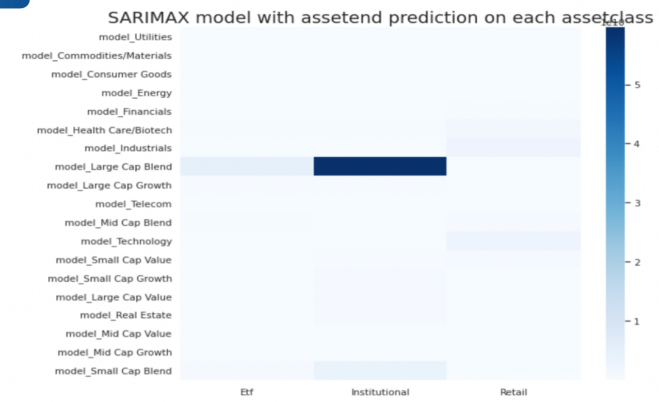
The first graph is the overall prediction trend compared with the actual test data, and the second and third graph is the trading signal prediction comparison between test and predicted dataset. Also, the AUC score will be displayed at the end of the function, for example, 69% in this case.

Evaluation:

As a summary, we plotted the heatmap for all asset classes in each of the three financial dataset with MSE and AUC metrics. We can see the model predicts AssetsEnd well for almost all asset classes in three dataset except the Large Cap Growth class in institutional dataset. And the model predicts tradeable signals well for almost all asset classes in three dataset except the Utilities, Commodities/Materials and Telecom class.

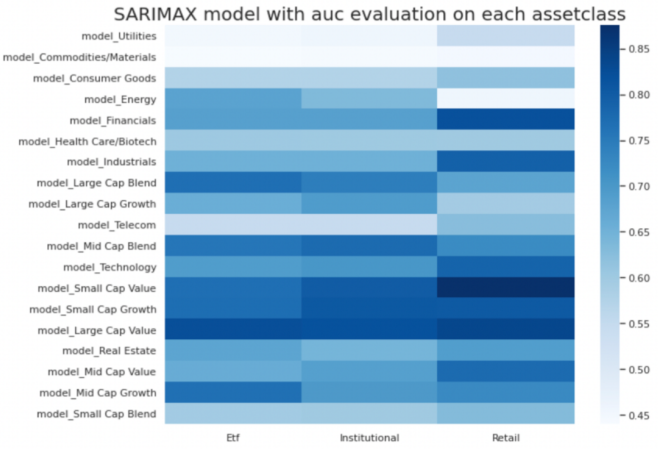
MSE Metric

	Etf	Institutional	Retail
model_Utilities	1.012354e+08	9.688173e+07	2.671870e+07
model_Commodities/Materials	1.726491e+08	1.777422e+08	1.506078e+08
model_Consumer Goods	3.050750e+07	5.274480e+07	3.482008e+06
model_Energy	9.074193e+07	2.618927e+07	1.295298e+07
model_Financials	3.170412e+07	2.714843e+07	4.386883e+08
model_Health Care/Biotech	4.328380e+08	4.397761e+08	1.860628e+09
model_Industrials	1.203487e+07	2.505699e+07	2.886196e+09
model_Large Cap Blend	4.957666e+09	5.961074e+10	1.519108e+07
model_Large Cap Growth	4.597751e+08	3.805982e+08	5.239626e+02
model_Telecom	7.881575e+04	1.513959e+05	9.743626e+06
model_Mid Cap Blend	2.732450e+08	2.310788e+08	8.586540e+08
model_Technology	7.983608e+06	4.582365e+07	2.820192e+09
model_Small Cap Value	3.840652e+07	4.176206e+08	8.553254e+08
model_Small Cap Growth	4.685072e+06	8.650262e+08	7.967155e+06
model_Large Cap Value	9.706309e+07	7.840565e+08	1.976894e+07
model_Real Estate	8.959693e+07	7.334740e+08	1.338696e+08
model_Mid Cap Value	1.494160e+07	2.159457e+07	1.947389e+05
model_Mid Cap Growth	7.571073e+06	7.878874e+07	1.721652e+04
model_Small Cap Blend	7.389203e+08	3.662471e+09	4.804297e+03



AUC Metric

	Etf	Institutional	Retail
model_Utilities	0.450820	0.458333	0.548513
model_Commodities/Materials	0.440000	0.444444	0.450000
model_Consumer Goods	0.575368	0.575368	0.620098
model_Energy	0.678922	0.634545	0.458333
model_Financials	0.684351	0.683817	0.820312
model_Health Care/Biotech	0.601852	0.600000	0.598214
model_Industrials	0.653639	0.653639	0.793103
model_Large Cap Blend	0.768293	0.744444	0.677419
model_Large Cap Growth	0.658570	0.692577	0.595987
model_Telecom	0.547181	0.547181	0.625632
model_Mid Cap Blend	0.756250	0.777027	0.721677
model_Technology	0.691106	0.702156	0.787500
model_Small Cap Value	0.768293	0.800956	0.875214
model_Small Cap Growth	0.770018	0.805735	0.804087
model_Large Cap Value	0.822874	0.818732	0.837209
model_Real Estate	0.675595	0.644720	0.688419
model_Mid Cap Value	0.662295	0.685152	0.778091
model_Mid Cap Growth	0.767875	0.698335	0.724405
model_Small Cap Blend	0.597561	0.598765	0.630824



Business Impact:

As an application of this model to the business world, the funding manager can use the SARIMAX model function to plot the AssetEnd trend and trading signal plots for asset classes that he or she is interested to explore. After getting the prediction accuracy, the manager can then pass this information to the finance team to evaluate the risk and return for making different investment strategies.

Conclusion and Limitation:

As a conclusion, the SARIMAX model did well with prediction of 116 weeks for all three datasets, but it does not capture the Large Cap Blend industry that well compared with other industries we have. The model is better at predicting the “buy” signals than “sell” signals. The accuracy or AUC score of this method varies between different asset classes, but works the best for retail dataset compared with ETF and Institutional dataset, with an average of 60% accuracy. When making actual investment decisions, one should consider whether the seasonal factor is significant enough to act, and we can always change the parameter to fit different prediction intervals or situations. But the model only counts as a reference before making final investment decisions.

10. Double Exponential Smoothing - With deseasonalized data

Double Exponential Smoothing (DES) is a method for capturing the trend of a time series and make predictions by past data. The underlying mathematics are shown in below's function:

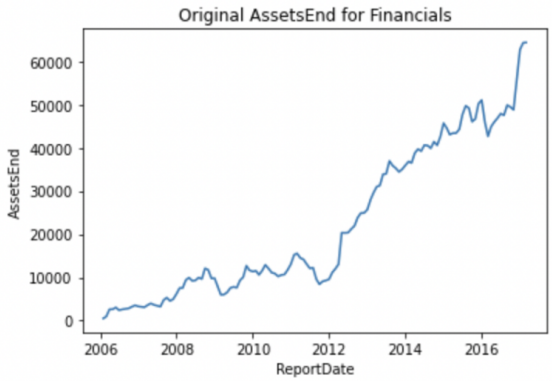
$$F_t = \alpha A_{t-1} + (1-\alpha) FIT_{t-1}$$
$$T_t = \beta (F_t - F_{t-1}) + (1-\beta)(T_{t-1})$$
$$FIT_{t+k-1} = F_t + kT_t$$

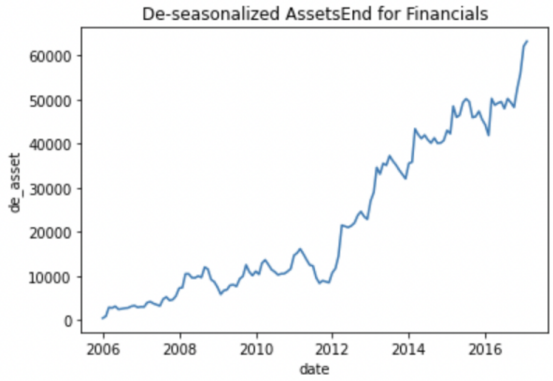
FIT : Forecast including trend
α : Base smoothing constant
β : Trend smoothing constant
k : Number of periods to forecast in the future
A : Actual observation

The two parameters (α and β) are the ones we need to tune for best performance.

Data Preprocessing:

Before implementing DES, we needed to deseasonalize the time series to remove the impact of seasonality. Our raw data was weekly data, so we first took the monthly average to resample them. We took each month as a season and calculated the corresponding seasonal factor and then divided the original data with their seasonal factor.

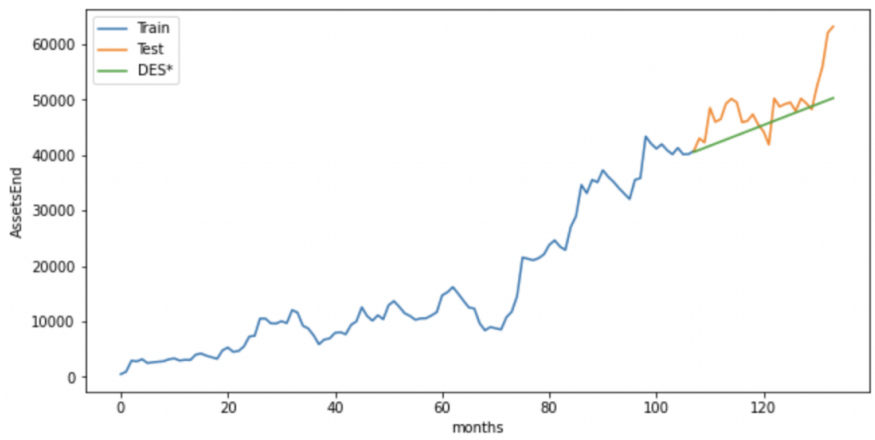




Comparing the two time series, we reckoned that deseasonalized one reflected the trend of the original series while removing seasonality.

Training and Predicting:

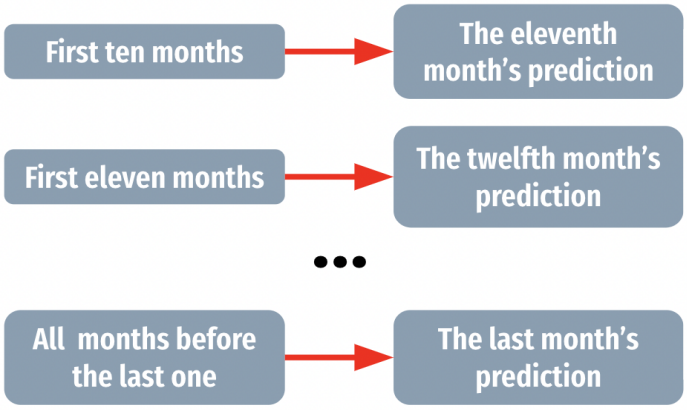
Train-test ratio: 80-20
Target variable: AssetEnd (deseasonalized)
Burn-in period: 10 months
Grid search: Try 0-1.0 with 0.05 step for both α and β to find the best combination



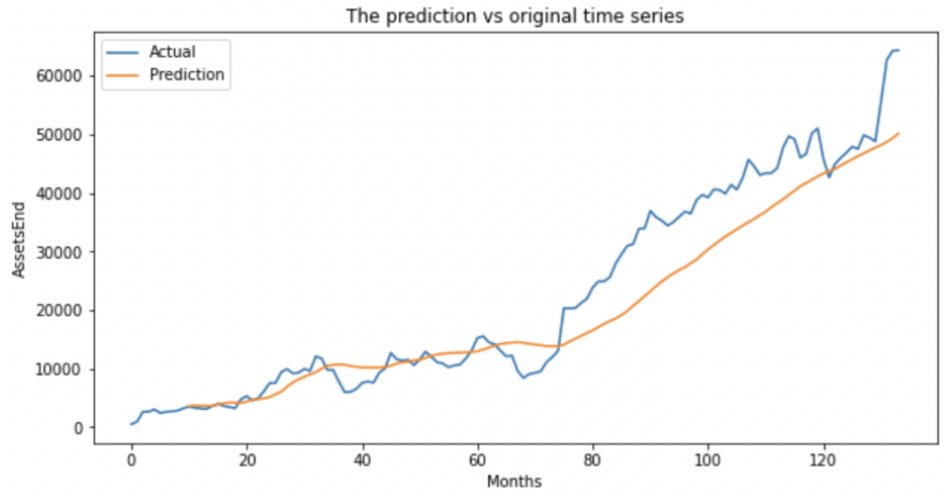
The test MSE for this example is 23931604.92, which is about the square of 4892. Since most observations in the test set are between 50000 to 60000, this is a satisfying result. However, we cannot implement this pipeline for identifying tradable signals because it's not giving us the information of the train set.

Extrapolate to the entire time series:

We thought of a method to extrapolate DES to the entire time series that can make predictions across the period. Like before, we still assumed a burn-in period of 10 months. Starting from the eleventh month, we used the data of the first ten months for prediction. Then we used the data from the first eleven months to predict the twelfth month. This method was repeating the process of DES across the whole period by adding new data for training.



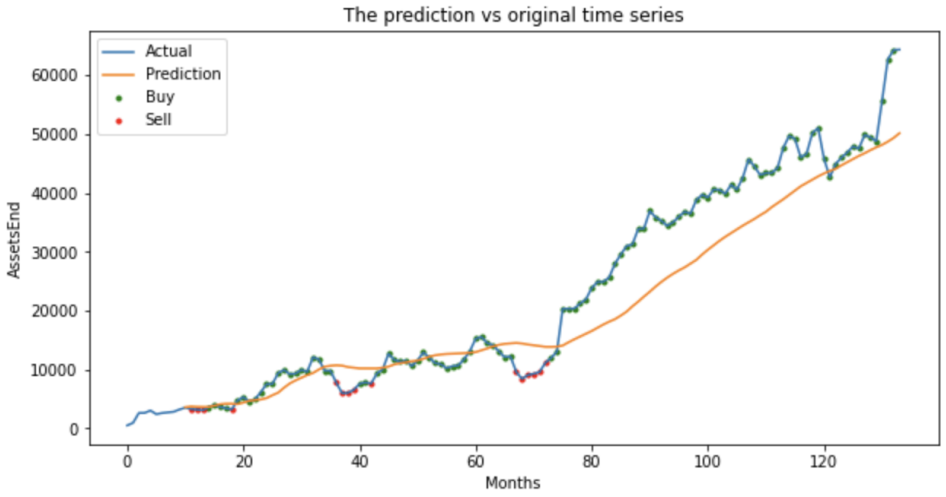
In the end, we received a smooth prediction of the time series which covers the entire period except for the first ten months.



Identify tradable signals:

We had the original time series and the smooth prediction of it. We could identify the tradable signals based on this information. The idea is to compare the prediction for next month and the prediction of this month. If the prediction for next month is higher than the previous, we identified that we should "buy" at this month since the value would go higher. If it's the opposite, we identified that we should "sell". We performed the same calculation on the original series to receive the "actual signal". If a fund manager was trading by this "actual signal", they would always make profit.

We plotted green dots as "buy" signals and red dots as "sell" signals.



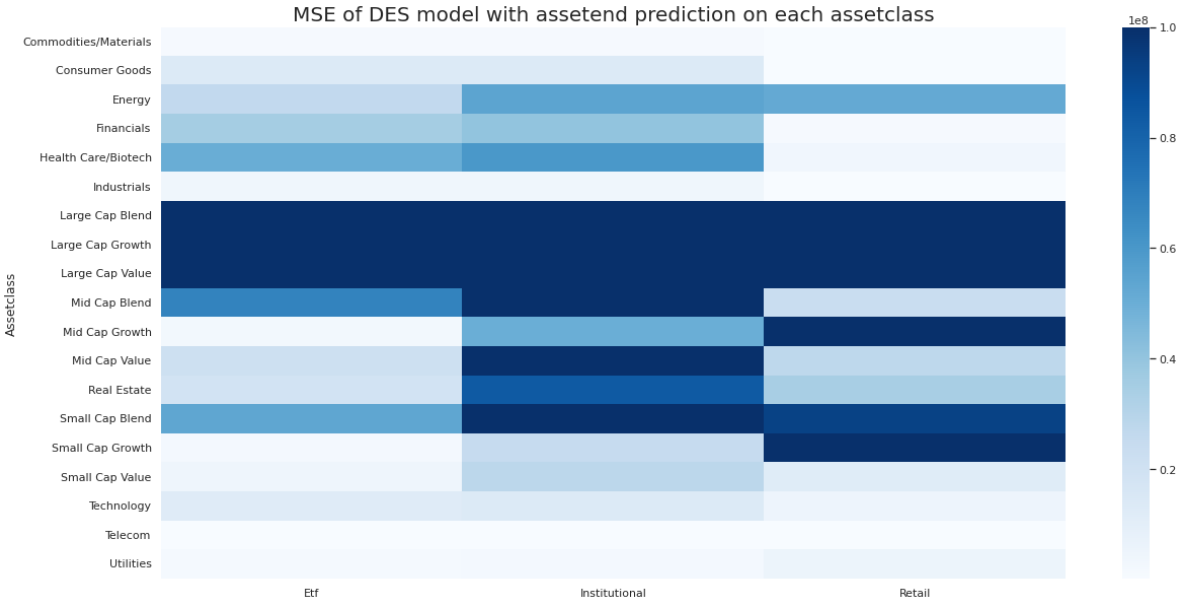
Evaluation:

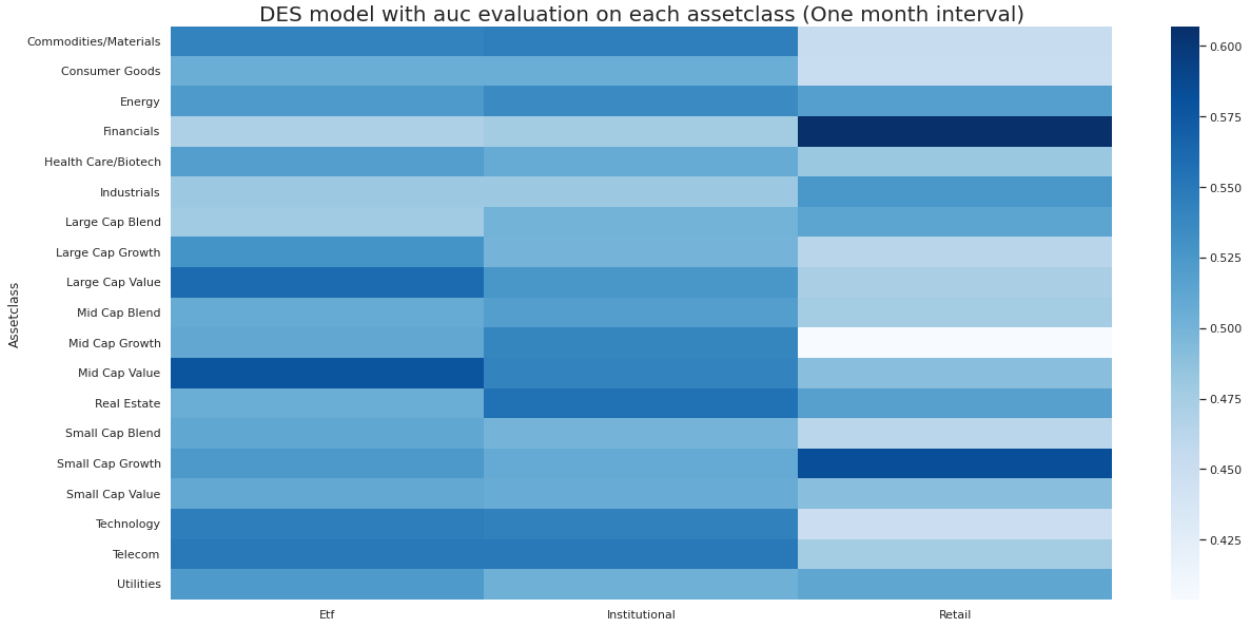
Since DES is more suitable for capturing the overall trend rather than single prediction, we could receive better results if we enlarged the interval by which we identified the signals. We tried three-months intervals and six-months intervals after the single month one.

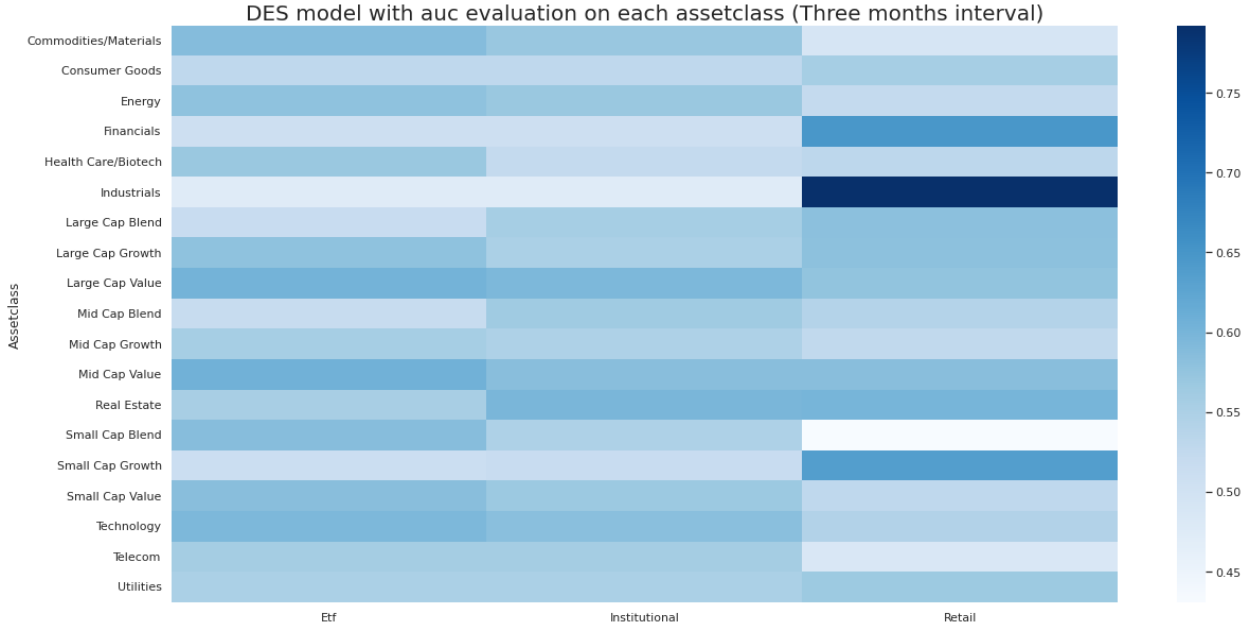
We built a function that takes in "asset class", "dataset" and "target column (AssetsEnd/PorfolioChangePct)" and outputs all corresponding results like below:

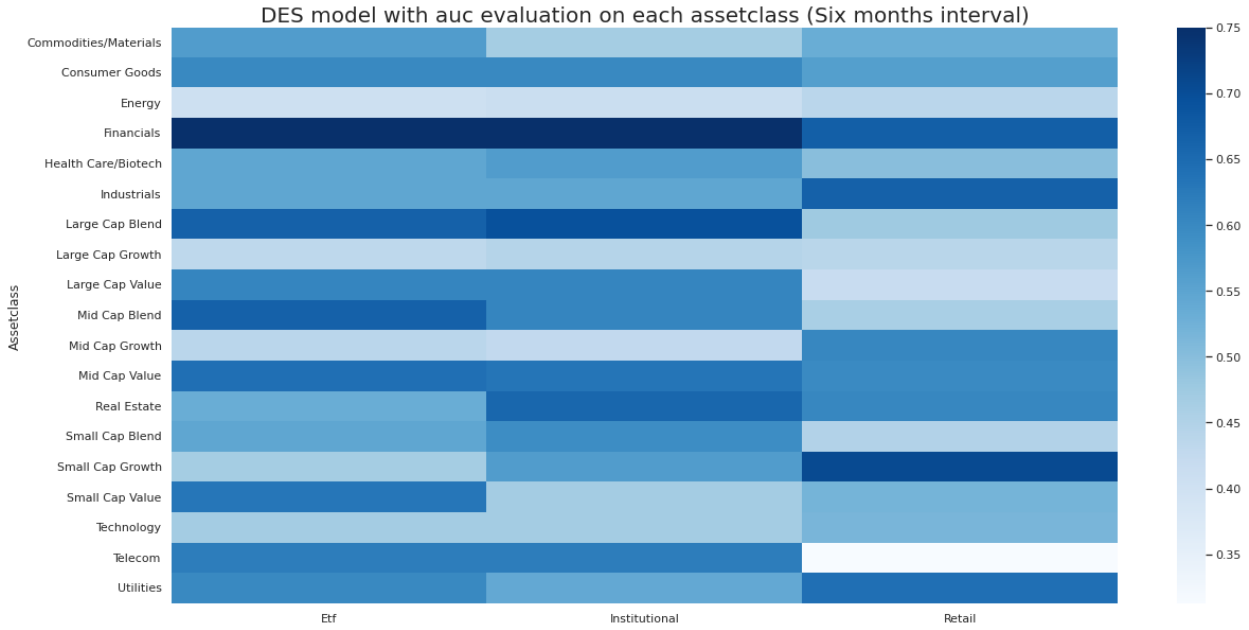
- Plot of original time series
- Plot of deseasonalized time series
- Plot of original time series and smooth prediction along with tradable signals (one, three, six months interval)
- MSE of original time series and smooth prediction
- Accuracy (one, three, six months interval)
- AUC (one, three, six months interval)
- True positive (one, three, six months interval)
- False negative (one, three, six months interval)
- True negative (one, three, six months interval)
- False positive (one, three, six months interval)

After running the function through all asset classes of the three datasets, we received the following results shown as heatmap.




DES model with auc evaluation on each assetclass (One month interval)
Assetclass
Commodities/Materials
Consumer Goods
Energy
Financials
Health Care/Biotech
Industrials
Large Cap Blend
Large Cap Growth
Large Cap Value
Mid Cap Blend
Mid Cap Growth
Mid Cap Value
Real Estate
Small Cap Blend
Small Cap Growth
Small Cap Value
Technology
Telecom
Utilities
Etf
Institutional
Retail
0.600
0.575
0.550
0.525
0.500
0.475
0.450
0.425


DES model with auc evaluation on each assetclass (Three months interval)
Assetclass
Commodities/Materials
Consumer Goods
Energy
Financials
Health Care/Biotech
Industrials
Large Cap Blend
Large Cap Growth
Large Cap Value
Mid Cap Blend
Mid Cap Growth
Mid Cap Value
Real Estate
Small Cap Blend
Small Cap Growth
Small Cap Value
Technology
Telecom
Utilities
Etf
Institutional
Retail
0.75
0.70
0.65
0.60
0.55
0.50
0.45



Business Impact:

The function built by DES would allow the team to conveniently choose the targeted asset classes from a certain dataset and provide the tradable signals. When implementing it in a real trading scenario, clients only need to update the dataset monthly and run the function again for reference. However, DES is only giving a smooth prediction, one should always consider whether the fluctuation is significant enough to action when making trading decisions.

Conclusion and Limitation:

- Double Exponential smoothing may not do well with prediction of a single month's value, but it's good at capturing the overall trend. So it could be used for identifying tradable signals of longer period.
- The accuracy or AUC score of this method varies between different asset classes, because they all have different trends during the 10 years between 2006 to 2017.
- Since most of the asset classes have an upward trend between 2006 to 2017, the model we have now does much better at predicting the "buy" signals than "sell" signals.
- We can refine the model if we have a larger scale of the data, preferably with more balanced "buy" and "sell" signals.
- By going over the market research, we could try more valid choices of time intervals besides one, three or six months.

11. Model Evaluation

Category	Classification		Time - Series		Neural Net	Forecast
Model	**Logistic Regression**	**VAR**	**ARIMA**	**SARIMAX**	**LSTM**	**DES**
Train/Test Ratio	80/20	Test on last 10 weeks	90/10	80/20	80/20	NA
Target Variable	Label	AssetEnd	AssetEnd	AssetEnd	Label	AssetEnd
Predictors	Flow FlowPct AssetEnd PortfolioChange%	Flow FlowPct AssetEnd PortfolioChange%	AssetEnd (Train/Test)	Flow FlowPct AssetEnd PortfolioChange% Moving Average 16/50 week	Flow FlowPct AssetEnd PortfolioChange%	AssetEnd
Evaluation Metric	AUC	RMSE	RMSE	MSE AUC	Accuracy	Accuracy/AUC /MSE

After evaluating all the models based on performance and application purpose, we chose SARIMAX and DES models as our final recommendation models. The SARIMAX model has performed well in various industries, and to a certain extent can be used as a guide for medium-term transactions. The DES model performed well for medium term transactions with a three month interval, but overall is better for long-term trend prediction. While the DES model provides the best consistent AUC overall, the SARIMAX model provides higher AUC on certain asset classes(e.g. Small Cap Value asset class).

12. Conclusion

In general, this project uses different models to fit transaction data with expected future market trends, and achieves the purpose of capturing the overall market trends and industry conditions.

From the model evaluation, the DES model has a better performance in capturing the overall market trend, and the SARIMAX model has a better performance in analyzing the market trends of various asset classes. In some specific asset classes, the SARIMAX model can even achieve an accuracy of about 80%, which can be used as a reference for investment.

This project still has one major limitation.

All the models described above use only market transaction data, the information contained in these data does not support us to conduct very accurate long-term market analysis. Therefore, it is necessary to expand the size of the dataset and introduce external data.

There are two directions for future improvement.

First of all, many external public data such as unemployment rate, interest rate, etc. are important references. Introducing these data can be used as a direction for future improvement.

Second, although this report briefly introduces the idea of applying the CAPM model to separate the rate of return affected by systemic risk, in practice, the model does not perform well, so this part is not presented in the report. Further improving this analysis path is also one of the ways to improve.

13. Documentation

13.1 Code Repository Location:

https://github.com/LinhTo11/Capstone

13.2 Python Tools

- pandas
- numpy
- matplotlib.pyplot
- seaborn
- LinearRegression
- SARIMAX
- statsmodels
- Scikit-learn

14 Promotional Purposes

We definitely learned a lot about data analysis of financial data and corresponding model building throughout the process of the capstone project. Therefore, we would like to include this valuable experience in the resume marked as *"Capstone Project".*

We plan on starting with an introduction of the datasets along with the goal we were trying to reach, which is defining and identifying the tradable signals. Next step we would like to briefly go through the EDA to present the features we learnt about the data during the process with plots and statistical information. Finally, we want to dig into all the models we tried and elaborate on the two models we systematically explored and finalized.

Inside any promotional material or content, we will only include the name of the sponsor company, faculty advisor or business advisor if we are given explicit permission.