PyTrader

Using Quantitative Measures to Trade Algorithmically

Outline

- Why use algorithmic trading
- The data
- Clustering stocks
- Technical Indicators
- Strategies
- Modeling stock commodities
- Backtesting strategies
- Making improvements

Algorithmic Trading

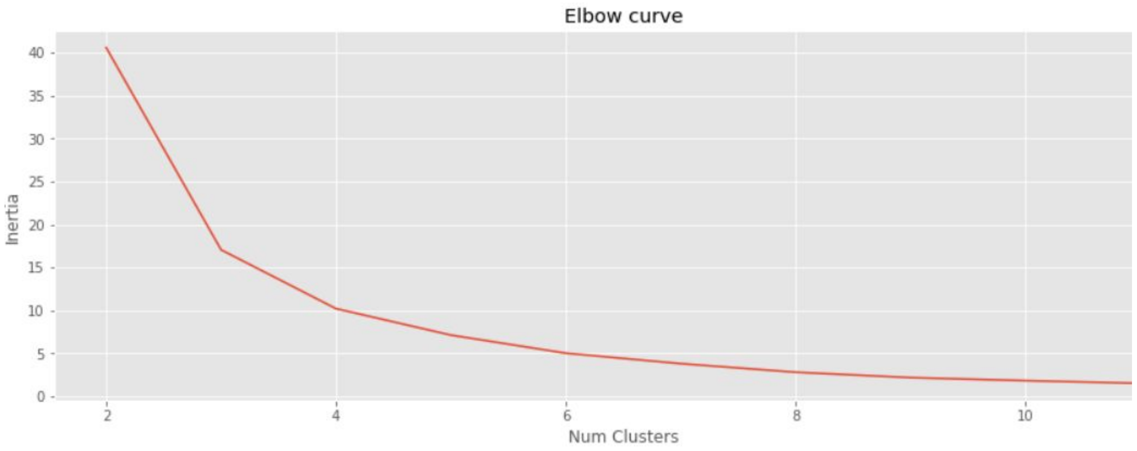
- What it is
  - Carrying out large amounts of trades based on pre-written rules calculated from multiple indicators such as price movement, volatility, volume, etc
- Why it’s important
  - The universe of stocks and trading is massive
  - Time is a very important consideration, especially in 24hr markets
  - Potential to remove human psychology as a hindrance
  - Allows us to better understand some trends in trading

The Data

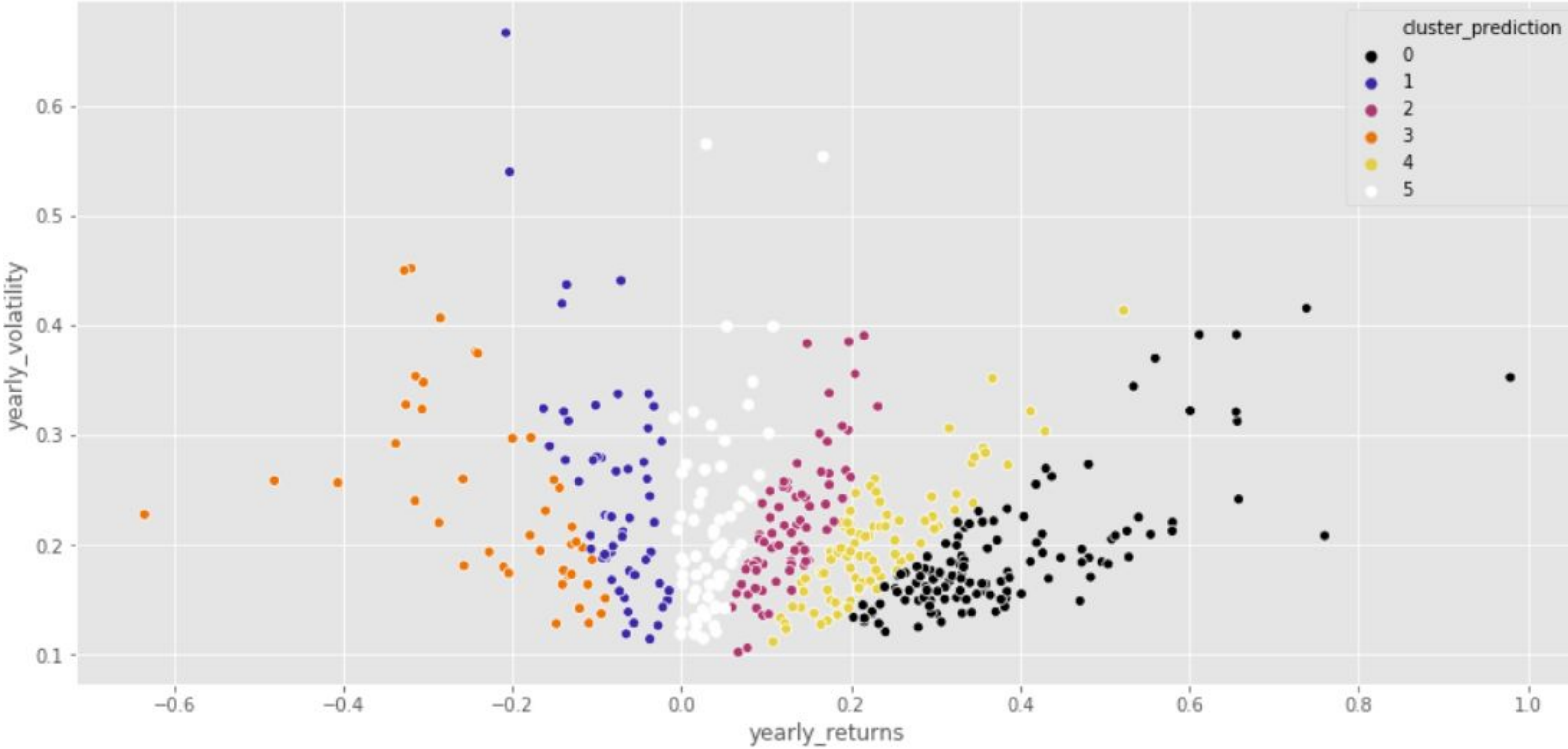
- Analyzing the US stock/equities market
  - Very granular data containing price information of various stocks
  - Data includes date, price (open, low, high, and close) and volume
  - Companies from the S&P 500 analyzed
- Date ranges from Jan 1st 2017 to Mar 25 2019 for strategy analysis
- Date ranges from Feb 2017 to Feb 2018 for clustering analysis
- Date ranges from Jan 2015 to March 2019 for LSTM
- Data can be gathered from Yahoo finance/Quandl/Kaggle
  - Depends on if the API is working
  - Kaggle data used for clustering
  - Quandle used for strategy analysis

Clustering stocks

- What factors can we use to cluster companies
  - Yearly returns and volatility
- Maybe see if we can use it to form a diverse portfolio
- KMeans to cluster and look at elbow plot of inertia to find n_clusters



Visualizing Clusters



Inspecting clusters by sector

Cluster 5 results

Utilities	15
Energy	9
Health Care	9
Consumer Discretionary	8
Financials	7
Industrials	6
Information Technology	4
Consumer Staples	3
Materials	3
Real Estate	2

Cluster 4 results

Industrials	23
Financials	19
Consumer Discretionary	15
Information Technology	9
Health Care	9
Materials	6
Energy	2
Utilities	2
Real Estate	2

Cluster 3 results

Real Estate	13
Consumer Discretionary	6
Energy	5
Consumer Staples	4
Communication Services	4
Utilities	3
Health Care	3
Industrials	2
Information Technology	1
Materials	1

Cluster 2 results

Financials	11
Information Technology	11
Industrials	10
Health Care	8
Materials	7
Consumer Staples	7
Consumer Discretionary	6
Real Estate	4
Energy	3
Communication Services	2
Utilities	2
Communication Services\n	1

Cluster 1 results

Consumer Staples	11
Consumer Discretionary	7
Communication Services	7
Energy	6
Industrials	5
Real Estate	5
Health Care	5
Information Technology	3
Utilities	3
Financials	3
Materials	1

Cluster 0 results

Information Technology	24
Health Care	21
Financials	21
Industrials	18
Consumer Discretionary	17
Communication Services	5
Consumer Staples	5
Materials	4
Real Estate	2
Energy	1

Focus strategies on one stock

- Analyzing Apple’s data to test strategy performance
- 2 years of data will be analyzed for analysis and ARIMA modeling
- 5 years of data will be analyzed for LSTM as it needs a larger input

Quantitative Analysis

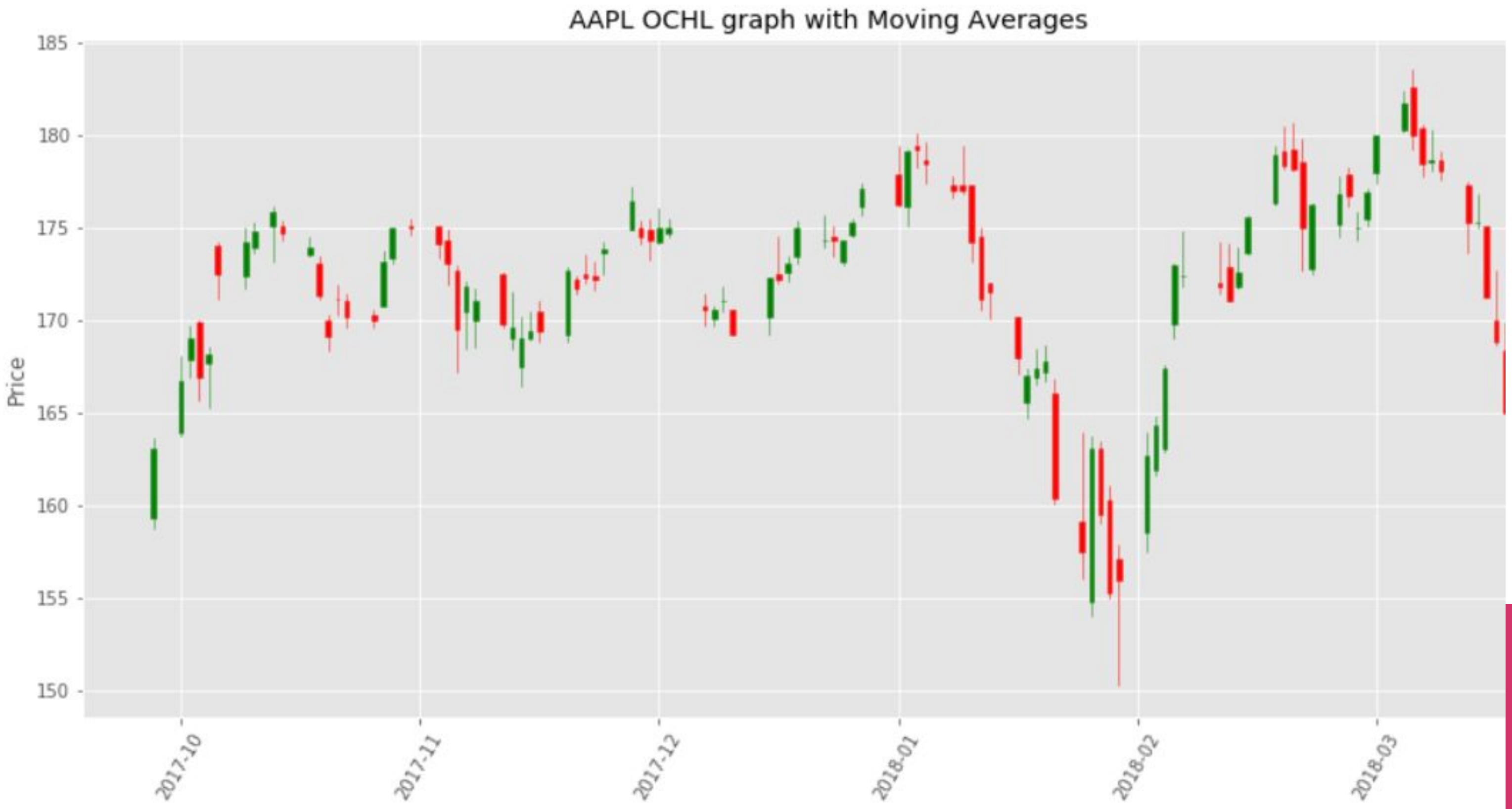
- Use of models and algorithms to model risk, understand behavior, and evaluate assets in investments
- Use of technical indicators to statistically quantify features for modeling
- Originally used through charting skills and visual analysis to understand trends by plotting technical indicators
    - Through computer science, these indicators were used to algorithmically carry out trades
- Technical indicators are any variables derived from price, volume, and time to describe various aspects of stock movement

Technical Indicators - 4 Major Categories

- Overlays
  - Trend indicators that overlap price movement
- Oscillators
  - Momentum indicators that show how fast prices moves
- Volume
  - Measures/confirms trend direction based on raw volume and its derivatives
- Volatility
  - Measures that show the deviation of a price over a given time period
- Amongst all the categories there are hundreds of indicators, so it’s best to choose anywhere from 1-3 indicators at a time in a strategy
  - Curse of dimensionality in noisy data

Visualizing the Data

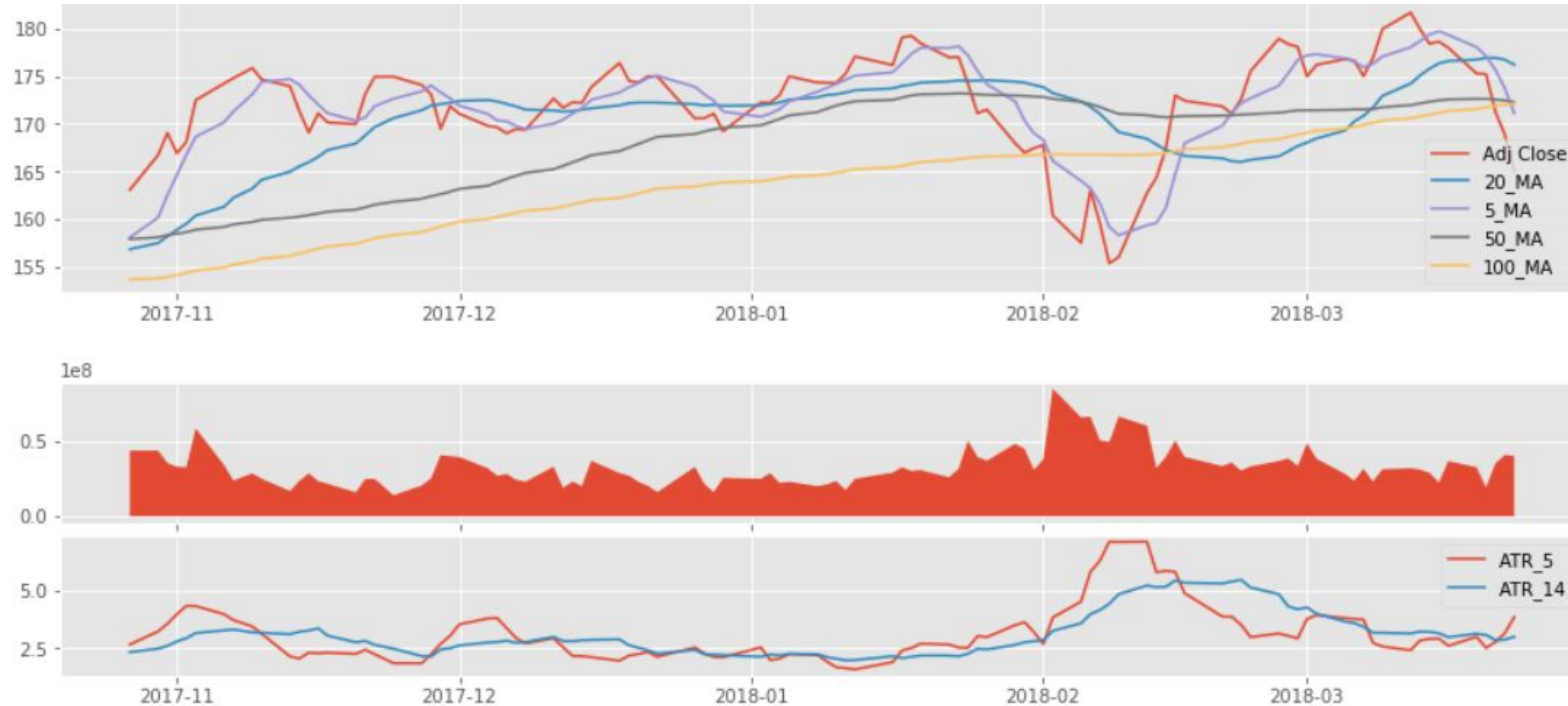
Candle Bar graphs



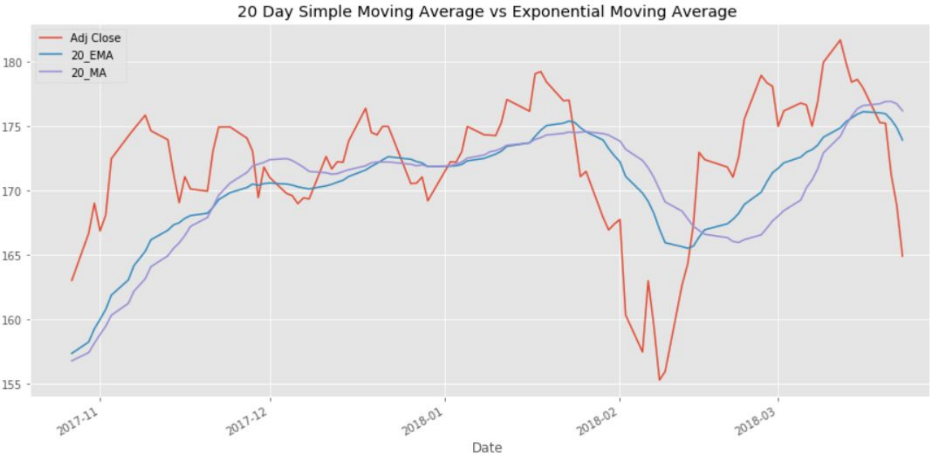
Overlays - Trend Indicators

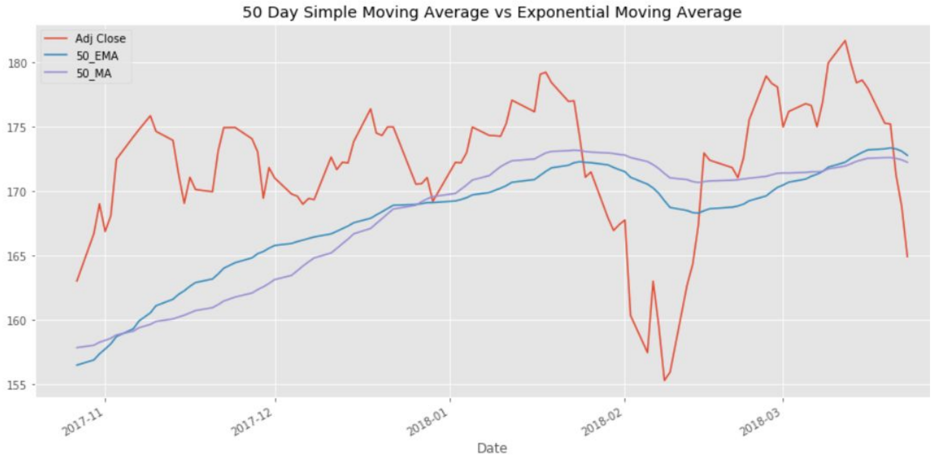
- Many variations to consider, all are some form of averaging the price for a given time period
  - Simple moving average (SMA), Exponential moving average (EMA), Kaufman Adaptive Moving Average (KAMA)
- Short windows follow the price trail more closely
- Longer windows can give the long term trend
- Crossover between short and longer windows allow us to create signals

Initial Indicator exploration



Comparing SMA to EMA





KAMA

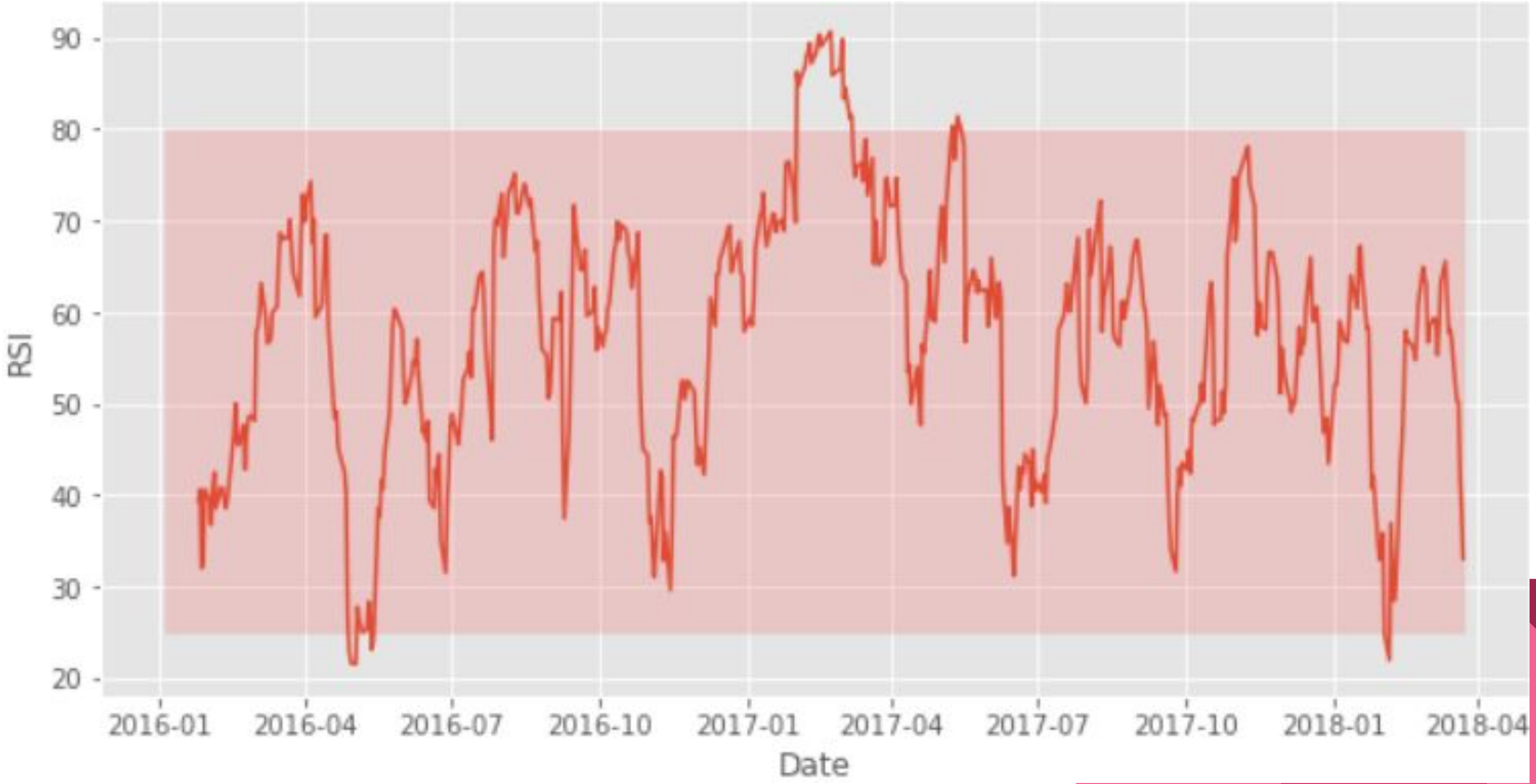
Adj Close
KAMA_10_2_30
KAMA_10_5_30
180
175
170
165
160
155
150
145
140
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11
2017-12
2018-01
2018-02
2018-03
2018-04
Date

Signals from KAMA

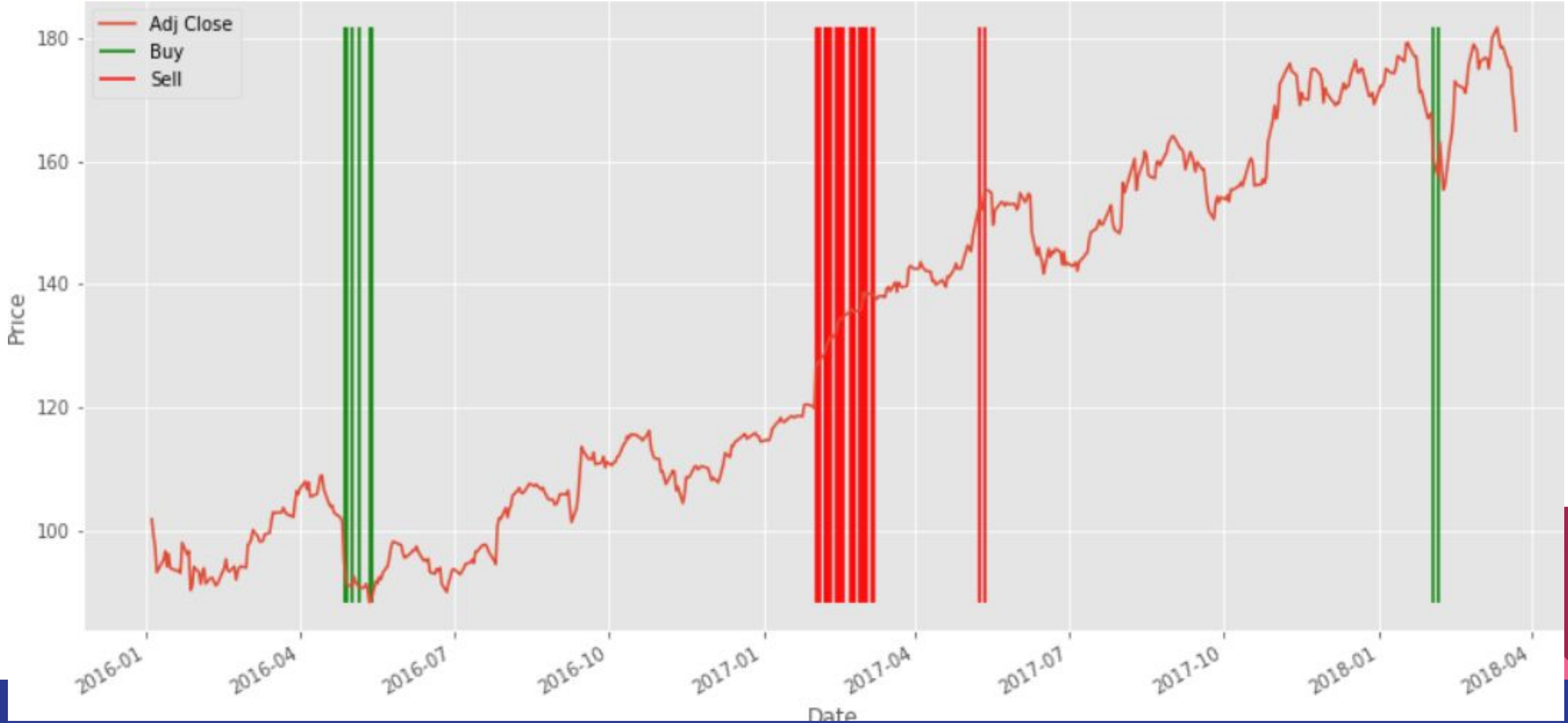
Oscillator - RSI

- RSI = 100 - 100/ (1 + RS)
  - RS is the average returns for a given time period, usually 14 days
- A momentum indicator that shows the strength of price fluctuations
- Ranges from 0 - 100
  - Usually, if RSI is above 70 then the stock is overbought and will likely lead to big sell off period
  - If RSI below 30, stock is undervalued which will likely lead to big buy period
- For this analysis, the upper range for RSI was set to 80 and lower range was set to 25

Analyzing signal creation potential



RSI Buy and Sell Signals



Volume signal

- From previous graphs, we see that volume and price correlate greatly
- On balance volume (OBV) was used, which indicates volume movement relative to buy and sell strength
- The if conditions are important
- Value of OBV doesn't matter as much as the conditions used to derive it

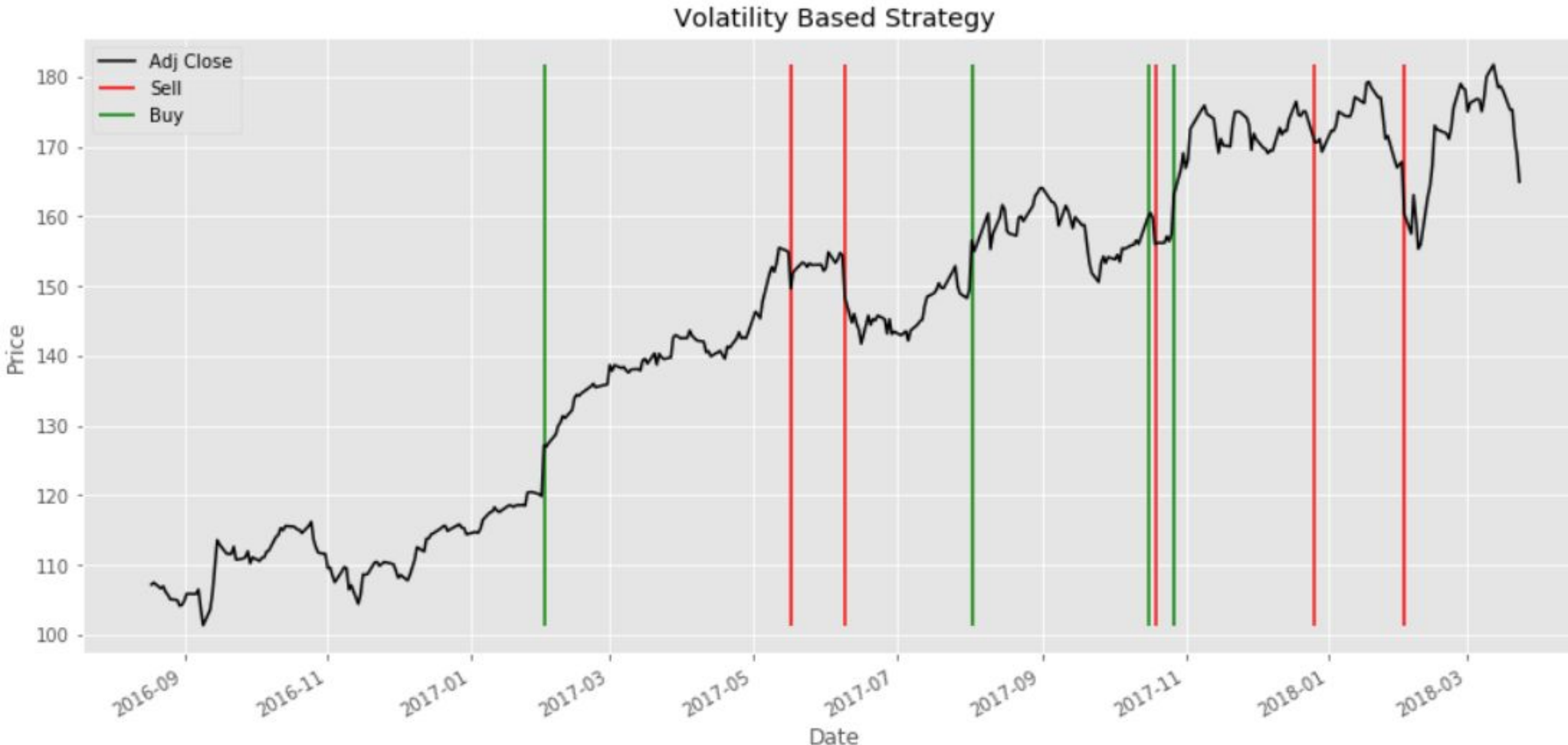
$$OBV = OBV_{prev} + \begin{cases} volume, & if close > close_{prev} \\ 0, & if close = close_{prev} \\ -volume, & if close < close_{prev} \end{cases}$$

OBV and Price

Volatility - ATR

- The volatility metric used is average true range (ATR)
    - Calculated as rolling average of the True Range
- True Range: max of the 3 values below for a day
    - High - Low
    - Current High - Previous Close (absolute value)
    - Current Low - Previous Close (absolute value)
- A rolling average of the volatility for a 5 day period and 14 day period were tested
- The 14 day period proved to be a good metric that was not as susceptible to sharp changes in price

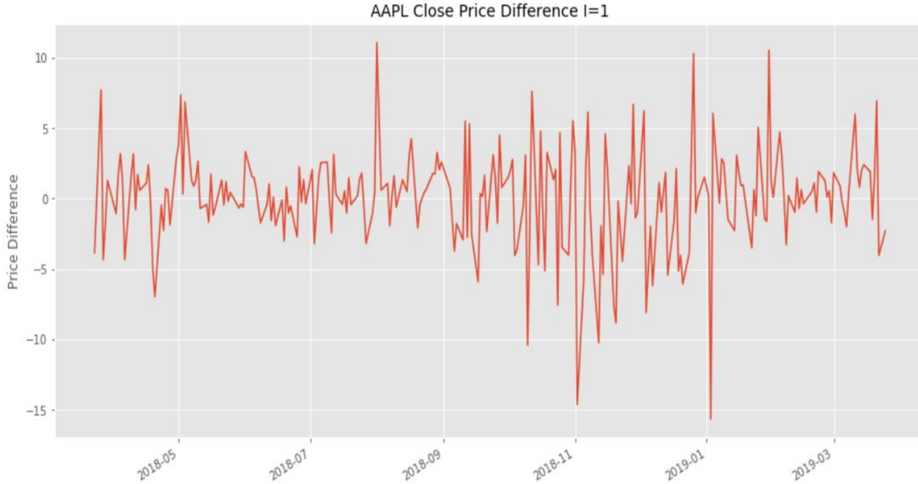
Volatility Signals



Beginning to model price

- Before beginning time series modeling
  - Data stationarity
- Data will need to be made stationary as stock data is very noisy
- Doesn't really have seasonality
- Stationarity will remove trends and allow the model to learn from a more normally distributed dataset

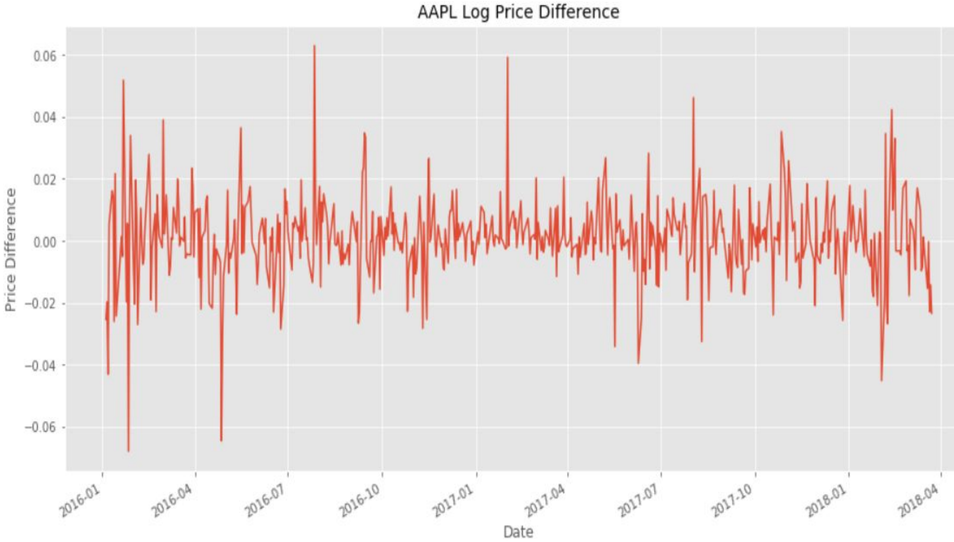
Beginning to model price




```
ADF Statistic: -1.924
p-value: 0.321
Critical Values:
        10%: -2.573
        1%: -3.458
        5%: -2.874
```

```
ADF Statistic: -15.996
p-value: 0.000
Critical Values:
        10%: -2.573
        1%: -3.457
        5%: -2.873
```

Log of the price to help with stationarity



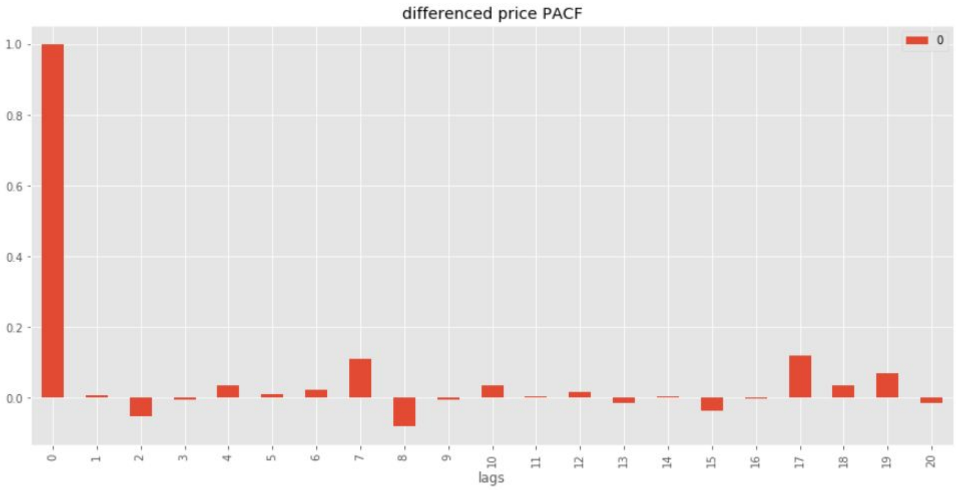

```
ADF Statistic: -23.415
p-value: 0.000
Critical Values:
        10%: -2.570
        1%: -3.442
        5%: -2.867
```

ARIMA Modeling

- Time series modeling technique that has 3 factors
    - Autoregressive component (AR) - effect of past values
    - Differencing component (I) - level of differencing to enforce stationarity
    - Moving average component (MA) - effect of previous error
- Requires that data is stationary and i.i.d
    - i.i.d. - independent and identically distributed
- Notation for order is ARIMA (p,d,q)
    - The p,d,q represent how far back in the should we look for each model component

Check ACF to inspect AR process

Checking PACF to pick parameters



Finding the best ARIMA model - regular price

ARIMA Model Results

Dep. Variable:	D.y	No. Observations:	454
Model:	ARIMA(4, 1, 2)	Log Likelihood	-346.749
Method:	css-mle	S.D. of innovations	0.513
Date:	Sun, 31 Mar 2019	AIC	709.497
Time:	18:55:46	BIC	742.442
Sample:	1	HQIC	722.477

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2236	0.099	2.261	0.024	0.030	0.417
ar.L1.D.y	1.3969	0.045	31.116	0.000	1.309	1.485
ar.L2.D.y	-1.2459	0.070	-17.907	0.000	-1.382	-1.110
ar.L3.D.y	0.8151	0.070	11.665	0.000	0.678	0.952
ar.L4.D.y	-0.3035	0.045	-6.746	0.000	-0.392	-0.215
ma.L1.D.y	-0.6107	0.011	-56.557	0.000	-0.632	-0.590
ma.L2.D.y	1.0000	0.029	34.741	0.000	0.944	1.056

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	0.0934	-1.2992j	1.3026	-0.2386
AR.2	0.0934	+1.2992j	1.3026	0.2386
AR.3	1.2496	-0.6170j	1.3936	-0.0730
AR.4	1.2496	+0.6170j	1.3936	0.0730
MA.1	0.3054	-0.9522j	1.0000	-0.2006
MA.2	0.3054	+0.9522j	1.0000	0.2006

- Using auto_arima, the it automatically finds the best model parameters by minimizing information criterion such as the AIC and BIC

Results


Price ARIMA (4,1,2)
train
actual
predictions
Close Price
240
220
200
180
160
140
120
2017-01
2017-04
2017-07
2017-10
2018-01
2018-04
2018-07
2018-10
2019-01
2019-04
Date

Best model params

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                    D.y   No. Observations:                  454
Model:                 ARIMA(1, 1, 4)   Log Likelihood                2044.189
Method:                       css-mle   S.D. of innovations              0.003
Date:                Sun, 31 Mar 2019   AIC                          -4074.379
Time:                        19:55:24   BIC                          -4045.552
Sample:                             1   HQIC                         -4063.021

==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0014      0.001      2.411      0.016       0.000       0.003
ar.L1.D.y     -0.0142      0.052     -0.271      0.786      -0.116       0.088
ma.L1.D.y      0.9769      0.028     35.443      0.000       0.923       1.031
ma.L2.D.y      0.9629      0.031     30.922      0.000       0.902       1.024
ma.L3.D.y      0.9274      0.057     16.149      0.000       0.815       1.040
ma.L4.D.y      0.9432      0.041     22.881      0.000       0.862       1.024
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1          -70.6319           +0.0000j           70.6319            0.5000
MA.1            0.3199           -0.9683j            1.0198           -0.1992
MA.2            0.3199           +0.9683j            1.0198            0.1992
MA.3           -0.8116           -0.6007j            1.0097           -0.3986
MA.4           -0.8116           +0.6007j            1.0097            0.3986
-----------------------------------------------------------------------------
```

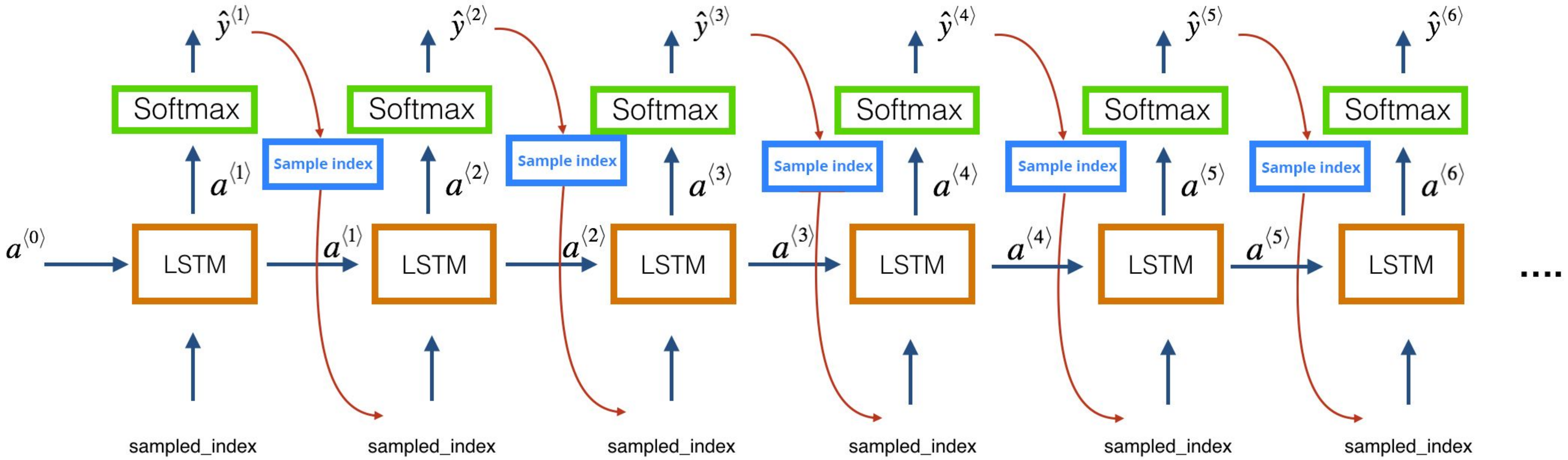
Results

Logged Price ARIMA (1,1,4)
train
actual
predictions
Close Price
240
220
200
180
160
140
120
2017-01
2017-04
2017-07
2017-10
2018-01
2018-04
2018-07
2018-10
2019-01
2019-04
Date

Using neural networks - LSTM

- Long Short Term Model (LSTM) is a type of recurrent neural network that differs from the basic feed-forward neural networks.
- Data can go back between layers and have a sense of memory when modeling
- Input format is generally (batch size, time steps, input dimension)

LSTM - How it works



LSTM input

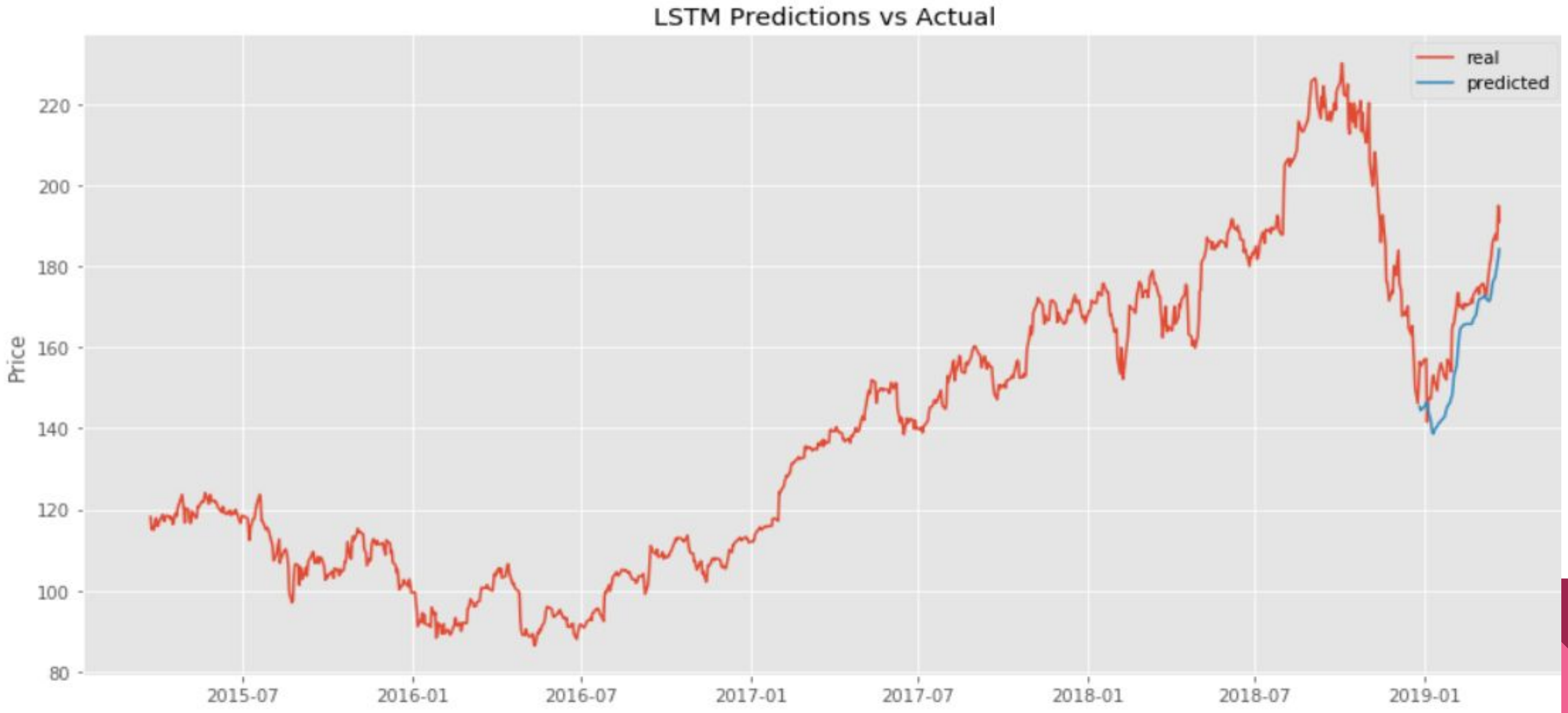
1	2	3	4	5
2	3	4	5	6
3	4	5	6	7
4	5	6	7	8
5	6	7	8	9

LSTM model creation

```
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 60, 50)            10400
_________________________________________________________________
dropout_1 (Dropout)          (None, 60, 50)            0
_________________________________________________________________
lstm_2 (LSTM)                (None, 60, 50)            20200
_________________________________________________________________
dropout_2 (Dropout)          (None, 60, 50)            0
_________________________________________________________________
flatten_1 (Flatten)          (None, 3000)              0
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 3001
=================================================================
Total params: 33,601
Trainable params: 33,601
Non-trainable params: 0
```

```
Epoch 1/10
887/887 [==============================] - 168s 189ms/step - loss: 0.0094
Epoch 2/10
887/887 [==============================] - 165s 186ms/step - loss: 0.0049
Epoch 3/10
887/887 [==============================] - 165s 186ms/step - loss: 0.0031
Epoch 4/10
887/887 [==============================] - 165s 186ms/step - loss: 0.0027
Epoch 5/10
887/887 [==============================] - 166s 187ms/step - loss: 0.0024
Epoch 6/10
887/887 [==============================] - 165s 186ms/step - loss: 0.0025
Epoch 7/10
887/887 [==============================] - 165s 186ms/step - loss: 0.0028
Epoch 8/10
887/887 [==============================] - 164s 185ms/step - loss: 0.0020s - - ETA: 2s
Epoch 9/10
887/887 [==============================] - 168s 189ms/step - loss: 0.0016
Epoch 10/10
887/887 [==============================] - 165s 187ms/step - loss: 0.0018
```

LSTM results



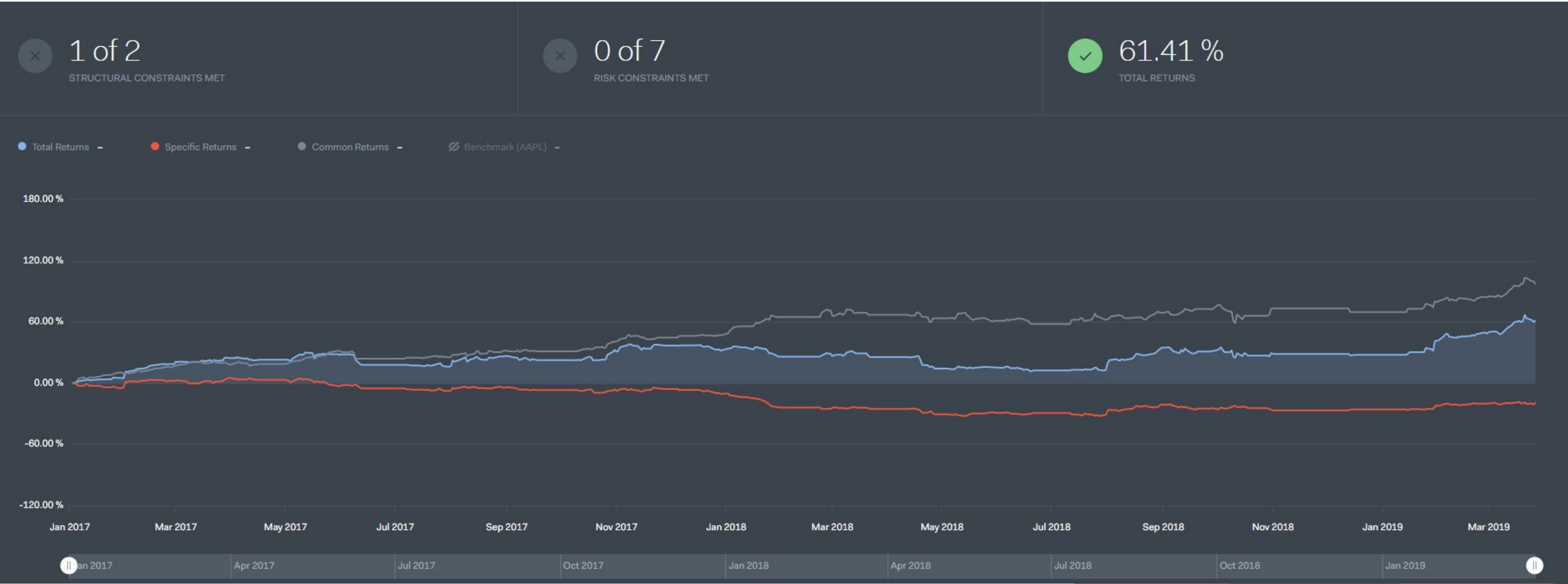
MSE for 60 day period: 85.55894235891884

Predicted vs Actual Signals

Backtesting

- Using strategy and results from modeling to test how well our strategies will work
- Survivorship bias
- Quantopian will be the platform to deploy our models on
  - Accounts for many factors such as time delays in ordering, changes in volume as companies increase their shares, and scheduling orders
  - Tracks multiple metrics to analyze our results
  - Test from Jan 1st 2017 to Mar 25th 2019
  - Base capital of $5000
- Primary metrics of concern will be total returns and sharpe ratio
  - Sharpe ratio is a measure of performance adjusted for risk
    - Ideal is to aim for ratio > 1

KAMA strategy



KAMA strategy

Total Returns 61.41 %

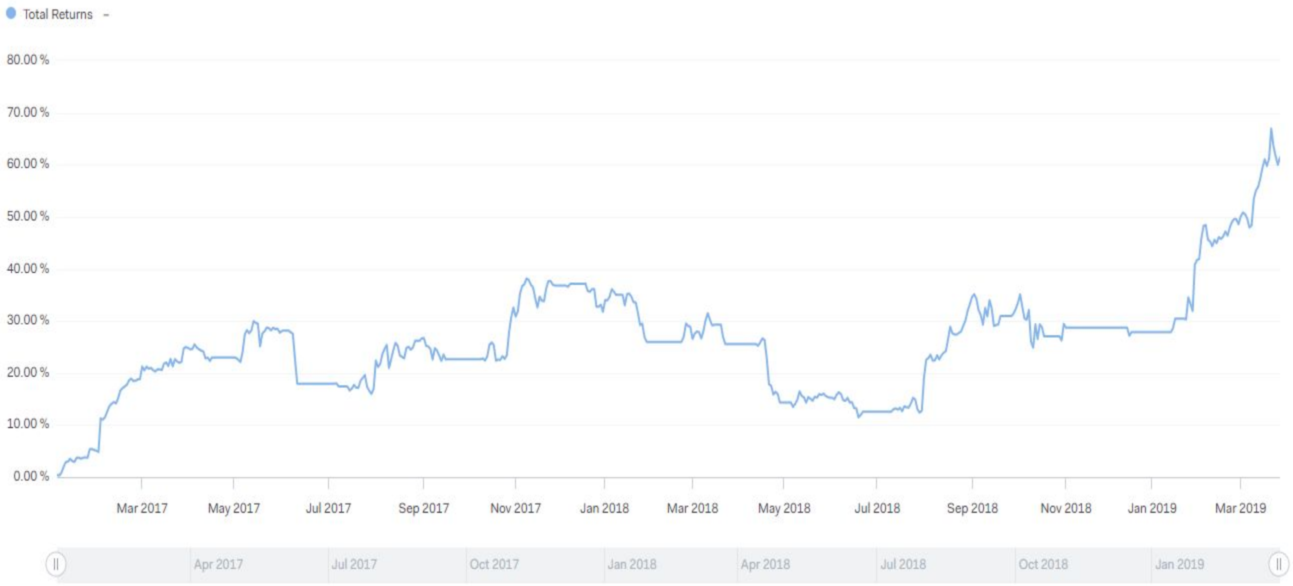
Specific Returns -19.09 %

Common Returns 97.04 %

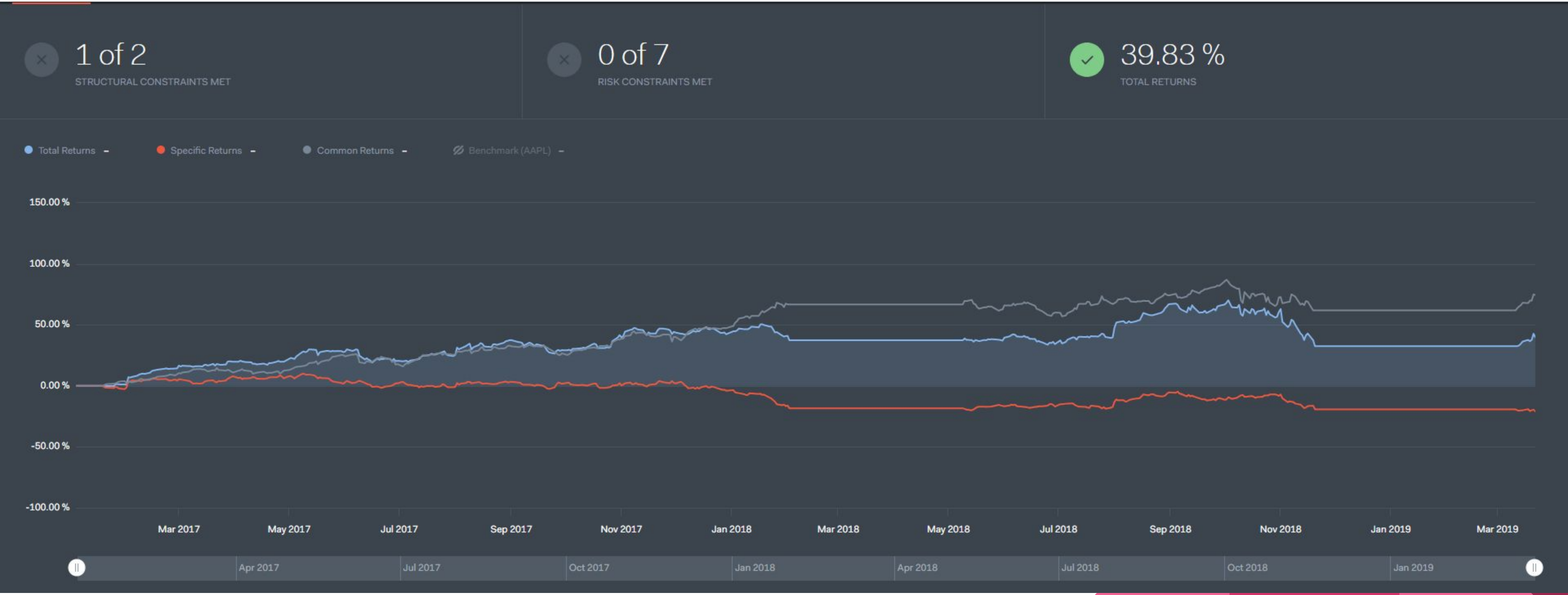
Sharpe 0.44

Max Drawdown -19.31 %

Volatility 0.16

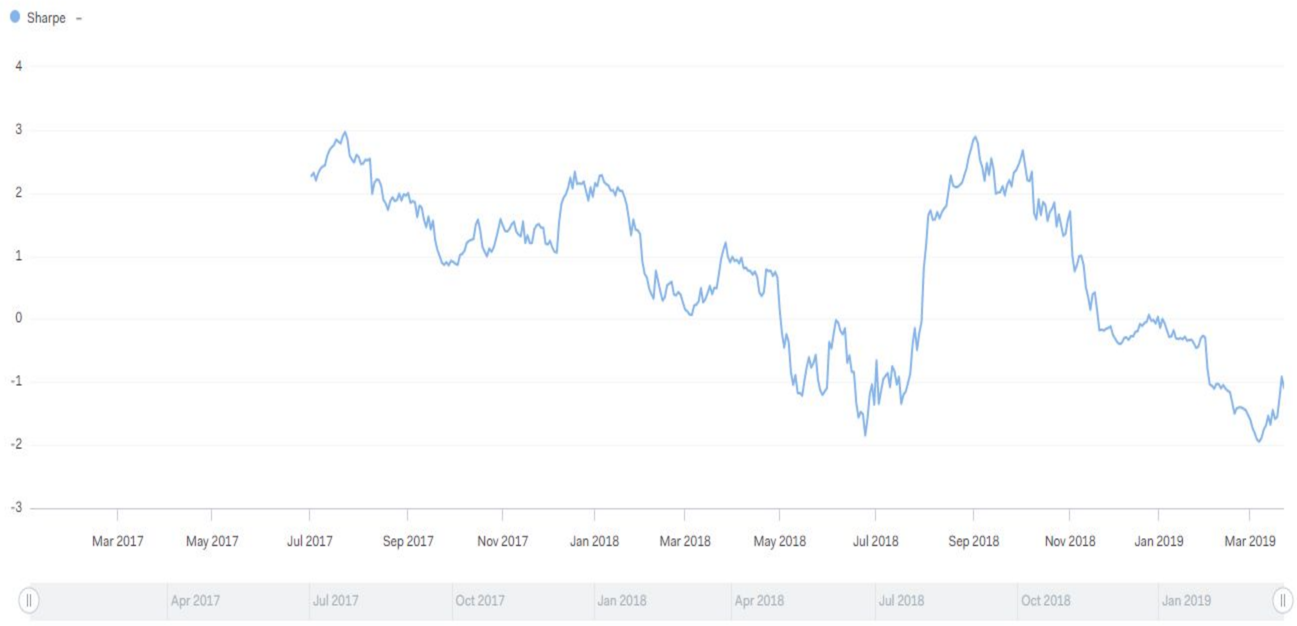


RSI strategy

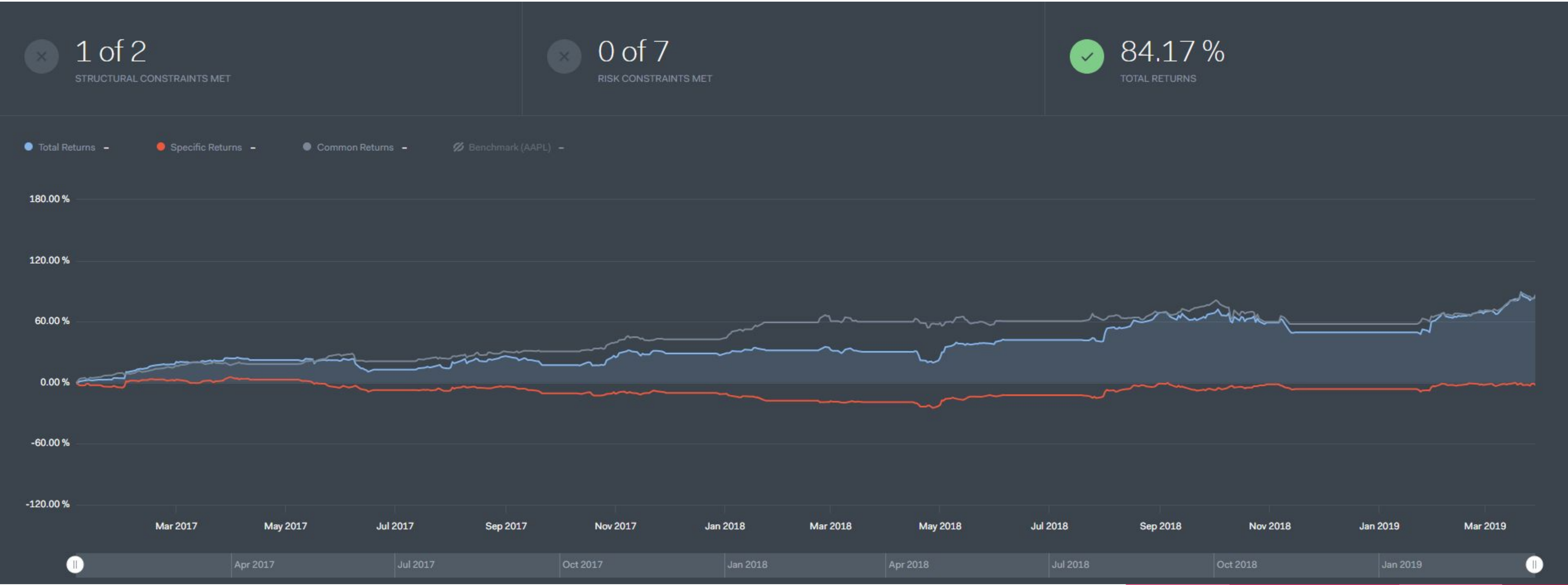


RSI strategy

Total Returns	39.83 %
Specific Returns	-20.95 %
Common Returns	74.51 %
Sharpe	0.75
Max Drawdown	-22.10 %
Volatility	0.18



OBV strategy



OBV strategy

Total Returns 84.17 %

Specific Returns -2.08 %

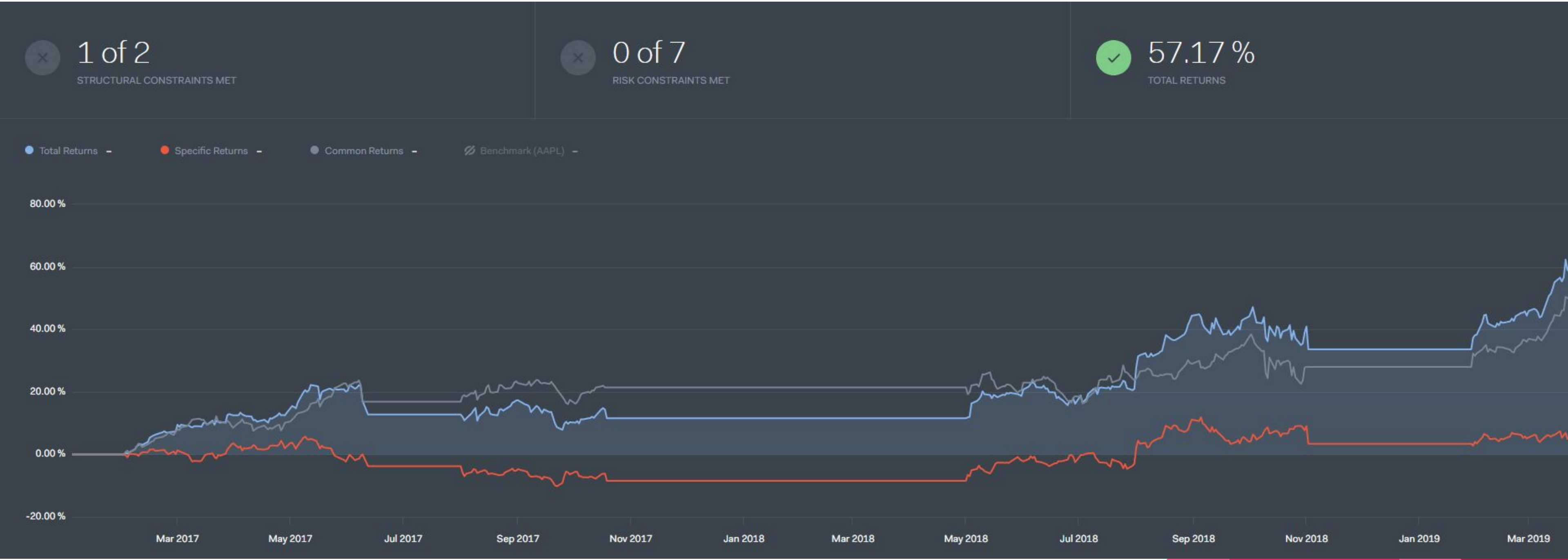
Common Returns 86.13 %

Sharpe 1.16

Max Drawdown -14.18 %

Volatility 0.17

Volatility strategy



Volatility strategy

Total Returns 57.17 %

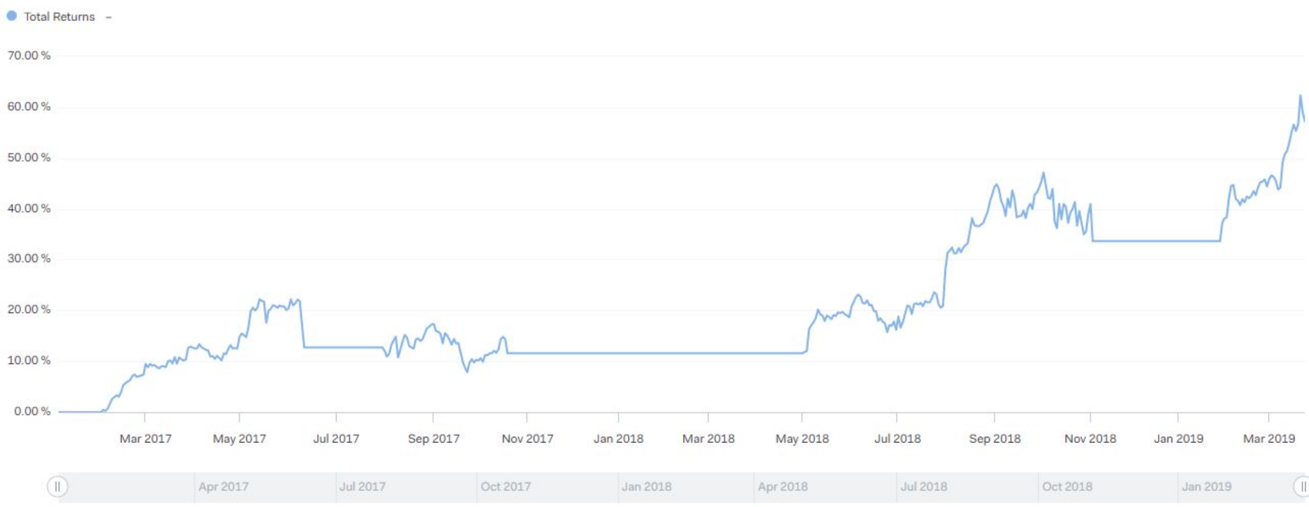
Specific Returns 5.28 %

Common Returns 47.81 %

Sharpe 0.97

Max Drawdown -11.73 %

Volatility 0.14



Best strategy so far

- The volatility strategy does will both in terms of performance and risk management
- Our base capital was low enough to offset the survivorship bias

Future improvements

- Combine multiple indicators to see if it performs better than a single strategy
- Use different strategies to enter and exit respectively
- Look into strategies and concerns for larger base capital
- Look into strategies implementing stock shorting
- Model ARIMA for returns as opposed to price and see if there's potential value
- Model volatility through GARCH and combine it with ARIMA
- Tune trading bot on quantopian to include multiple stocks accounting for diversity and sentiment
- The possibilities are endless in terms of strategy combinations