

Stocks Portfolio Design with *Numbers* and *Text* Dataset

Jingying (Jane) Bi¹, June 2018

Abstract

This paper designed a portfolio consisting of stocks which are “wonderful business at bargaining price with high momentum and great industrial leadership”. Firstly, I created a *raw portfolio* of 26 stocks selected through Portfolio 123. Then, I conducted sentiment analysis on 3,279,343 Reuters news with various statistical models to *predict* the performance of each single stock in *future*. With 21 promising stocks in gaining positive inflation-adjusted excess return, I further shortlisted 9 of them to make up of *ultimate portfolio*. These nine stocks have the least correlations among each other. The *actual* performance of the *ultimate portfolio* beat both the *raw portfolio* and the market benchmark in *future*.

Keywords: Stock ranking system, stock screening, textual analysis, sentiment analysis, big data, machine learning

¹ jingyingb@uchicago.edu | UCID: 12174556

1. Introduction

This paper created portfolio robustly beating S&P 500. Such portfolio consists of weighted-average stocks with abnormal returns, given the correlation of selected stocks is the most negative (or least positive). Note that I will only consider the *inflation-adjusted excess returns (IAER)* instead of nominal returns. This portfolio only takes “Long” positions. S&P500 is the benchmark.

My *text* dataset for later analysis covers Jan 1st, 2007 to Apr 29th, 2018. To smooth the whole analysis process, I made assumptions on several important dates. Suppose *today* is Nov 30th 2017. Before *today*, Sep 9th, 2015 divided the data into *training* and *testing* dataset. After *today* from Dec 1st, 2017 to Apr 30th, 2018 is defined as *future*, and correspondingly the *predicting* dataset. In short, I used the data from 2007-01-04 to 2015-09-09 to train the models, and the data from 2015-09-10 to 2017-11-30 to test the models. With the best performed models, I used data from 2017-12-01 to 2018-04-30 to predict stock’s *IAER*. Table 1.1 gives a summary of the time frame.

Table 1.1 *Training, testing and predicting dataset*

| Dataset | Date | Dataset size | Function |
|------------|-----------------------------|--------------|---|
| Training | 2007-01-04 ~ 2015-09-09 | 2178 | Training models |
| Testing | 2015-09-10 ~ 2017-11-30* | 545 | Test models’ performances and find the best model with highest out-of-sample R2 (OOS R2) |
| Predicting | 2017-12-01 ~ 2018-04-30 | 100 | Predict the excess inflation-adjusted returns (EIAR) of portfolio, and compare with benchmark S&P500. |

* Assuming *today* is 2017-11-30.

I firstly designed ranking system and screening on Portfolio123, with a simulation test. I kept adjusting the components, parameters, and rules until obtaining a well-performed portfolio in terms of annualized excess returns, Sharpe ratio, and overall winners percentage. I call it *raw portfolio*. There are 26 stocks in *raw portfolio*. Next, I worked on these 26 stocks, and create the *ultimate portfolio*. To make the *ultimate portfolio*, I firstly conducted textual and sentiment analysis to predict *IAER* with the *predicting* dataset, securing four stocks with highest probability of gaining positive *IAER* in future, putting another ten stocks on the candidate list. Among 4+n stocks (four secured stocks and n stock candidates), I selected the combination of stocks with the most negative (or least positive) correlation. There were nine stocks being shortlisted and they

made up of the *ultimate portfolio*. Last but not least, I used the actual dataset to backtest the performances of *raw portfolio*, *ultimate portfolio* and benchmark S&P500. The result shows that three portfolio had similar performance at the very beginning. After Feb 2018, the *ultimate portfolio* managed to yield much greater inflation-adjusted excess returns. *Raw portfolio* behaved better than S&P500 benchmark as well.

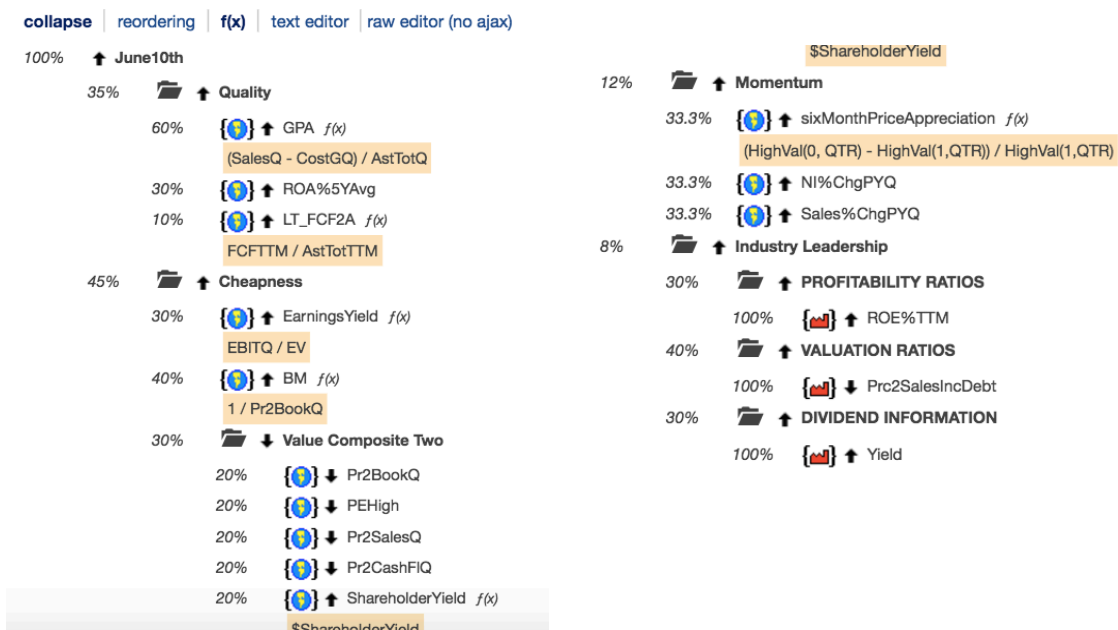
The rest of paper is organized as follows: Section 2 displays stock ranking, screening, and simulation on Portfolio123. Section 3 presents *text* data collection, cleaning and construction. Section 4 discusses model training, testing and predicting. Section 5 creates the *ultimate portfolio* and analyses the actual performances. Conclusion is in Section 6.

2. Stock Ranking and Screening on Portfolio123

2.1 Ranking System

My design principles followed the rationale of “buying wonderful business at bargaining price and with high momentum”. In addition, I took some more industry-specific factors into consideration, because I hope to select stocks not only with general promising performance but also being able to ace in its industry. Figure 2.1 shows the ranking system. The ranking system has four components, namely *Quality*, *Cheapness*, *Momentum* and *Industry Leadership*. Each component consists of multiple factors.

Figure 2.1: Ranking system



Most of the factors were selected based on my readings. For instance, under *Quality*, GPA is coined by Gray and Carlisle. GAP plays a similar role as Greenblatt's return on capital (ROC) which is an indicator of business quality. Return on asset (ROA) has been proposed by O'Shaughnessy as a factor while ranking stocks. Instead of using plain ROA, I used the 5-year average ROA, because I hope to select those stocks with robust performance in the mid-long term. Under *Cheapness*, I added the classic earnings yield proposed by Greenblatt as an indicator of bargaining price. Book-market ratio (BM) has been recommended by Gray and Carlisle as part of *Cheapness* component. Moreover, O'Shaughnessy coined their *Value Composite Two* to measure the cheapness of stocks. Their *Value Composite Two* consists of price-to-book (P/B), price-earnings (PE), price-to-sales (P/S), EBITDA/EV, price-to-cash-flow (P/CF) and shareholder yield. Hence, I imported this *Value Composite Two* into my ranking system. Note that I created the function of \$ShareholderYield in Portfolio123. Following O'Shaughnessy, I formulated \$ShareholderYield as:

$$\text{\$ShareholderYield} = (\text{DivPaidQ} + \text{EqIssuedQ} - \text{EqPurchQ} + \text{DbtLTIssuedQ} - \text{DbtLTReducedQ}) / \text{MktCap}$$

Figure 2.2 Basic settings of ranking system

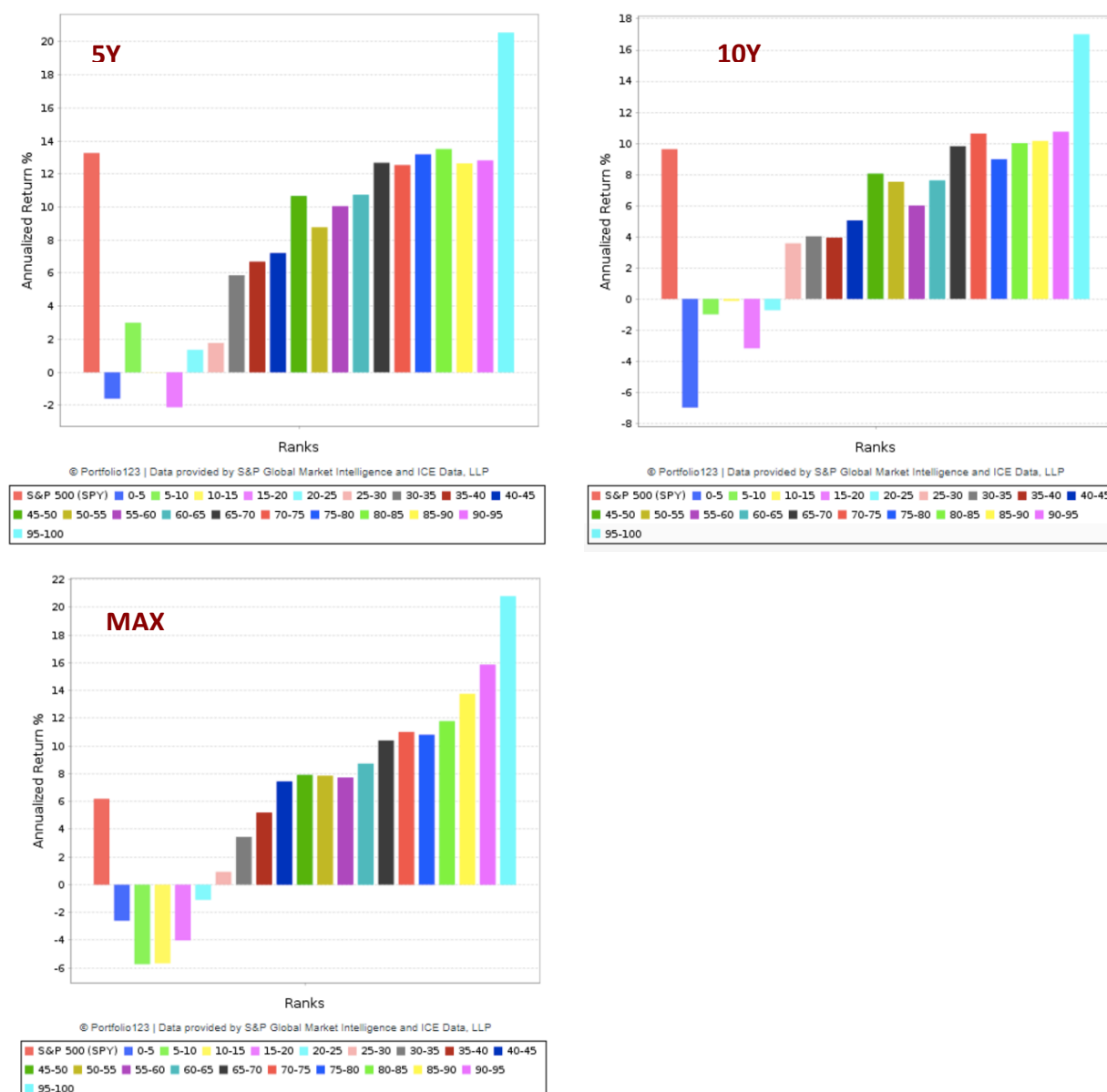
Historical Performance by Ranks Run

| | |
|-------------------------|---|
| Period | 1M 6M 1Y 2Y 5Y 10Y MAX |
| Rebalance Frequency | Every 13 Weeks |
| Default Ranking Method | Percentile NAs Neutral |
| Override Ranking Method | Percentile NAs Neutral |
| Benchmark | S&P 500 (SPY) |
| Universe | All Fundamentals - USA |
| Rank Buckets (2-200) | 20 |
| Slippage % | 0.5 |
| Transaction Type | <input checked="" type="radio"/> Long <input type="radio"/> Short |

Performance of this ranking system was evaluated based on 5Y, 10Y, and MAX period. The basic setting is displayed in Figure 2.2. Figure 2.3 shows the ranking performance of 5Y, 10Y and MAX. Three panels of Figure 2.3 all show a monotonic increase in the returns of bucket. The

monotonic upward sloping indicates that the system functions well in ranking stocks with different returns. Even better, the top one bucket in each panel always displays much higher return than the rest buckets and benchmark S&P500. This feature reinforced my trading strategy's performance, because I will only consider taking "Long" positions on stocks with abnormally high returns (i.e. top one bucket, rank > 95). Last but not least, checking different time periods (i.e. 5Y, 10Y, MAX) guarantees the robustness of my trading system.

Figure 2.3 Performance of ranking system

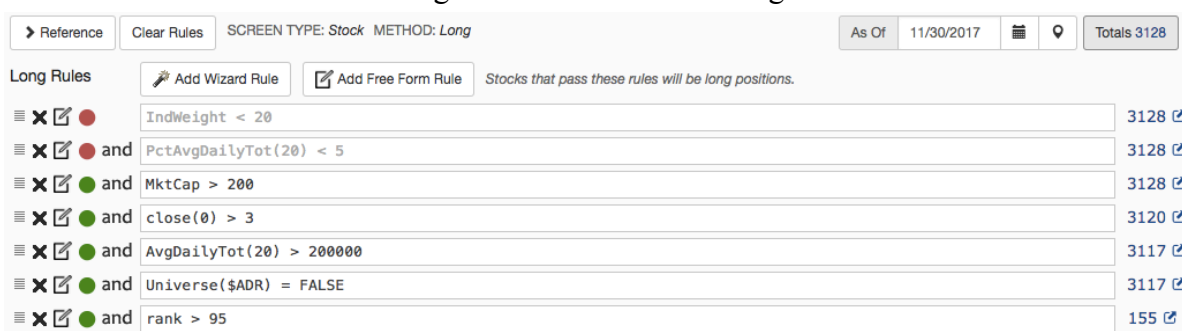


NOTE: Dividends are included. Transaction costs are not included. The positions in each 'bucket' are equally weighted regardless of market cap.

2.2 Screening

The screening contains five rules, as displayed in Figure 2.4. Setting limit on the market capitalization is because most outstanding market-beating returns are attributed to micro-companies. Although including them makes backtesting result more attractive, doing so makes the strategy less realistic, because these micro- stocks are usually not available in market for traders to buy or sell. Hence, I only consider companies with market cap larger than 200 million. A limit on close price filters out unhealthy stocks which successfully pass other rules but with extremely low prices. Average daily volume reflects a liquid and well-running business status. No foreign stocks are considered. Most importantly, according to the ranking system performance in Figure 2.3, I only buy stocks in the top one bucket whose ranking is above 95%. Figure 2.4 also gives number of stocks passing each rule. With ranking > 95 , 155 stocks will be considered as of 2017-11-30. (Remember *today* we are on Dec 30th 2017 :))

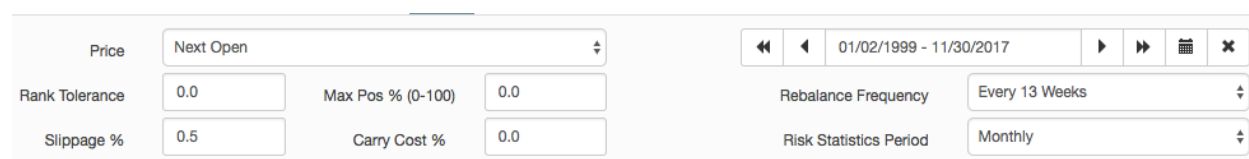
Figure 2.4 Rules of screening



| SCREEN TYPE: Stock METHOD: Long | | As Of | 11/30/2017 | Totals |
|-------------------------------------|------------------------------|-------|------------|--------|
| Long Rules | | | | |
| <input checked="" type="checkbox"/> | IndWeight < 20 | | | 3128 |
| <input checked="" type="checkbox"/> | and PctAvgDailyTot(20) < 5 | | | 3128 |
| <input checked="" type="checkbox"/> | and MktCap > 200 | | | 3128 |
| <input checked="" type="checkbox"/> | and close(0) > 3 | | | 3120 |
| <input checked="" type="checkbox"/> | and AvgDailyTot(20) > 200000 | | | 3117 |
| <input checked="" type="checkbox"/> | and Universe(\$ADR) = FALSE | | | 3117 |
| <input checked="" type="checkbox"/> | and rank > 95 | | | 155 |

The screening is followed by backtesting. Some basic setting before running backtesting is displayed in Figure 2.5. Slippage rate was 0.5%. Backtesting results are shown in Figure 2.6. the three panels tell that my trading strategy robustly beating the market in different time periods. It performed best in MAX followed by 10Y and 5Y. In addition to annualized returns, my portfolio performs well in terms of such key statistics as similar max drawdown with S&P500, higher Sharpe and Sortino ratio, and tolerable larger standard deviation. Table 2.1 gives a summary of statistics.

Figure 2.5 Basic settings of backtesting



| | | | | | |
|----------------|-----------|-------------------|-----|-------------------------|----------------|
| Price | Next Open | | | 01/02/1999 - 11/30/2017 | |
| Rank Tolerance | 0.0 | Max Pos % (0-100) | 0.0 | Rebalance Frequency | Every 13 Weeks |
| Slippage % | 0.5 | Carry Cost % | 0.0 | Risk Statistics Period | Monthly |

Figure 2.6 Backtesting results

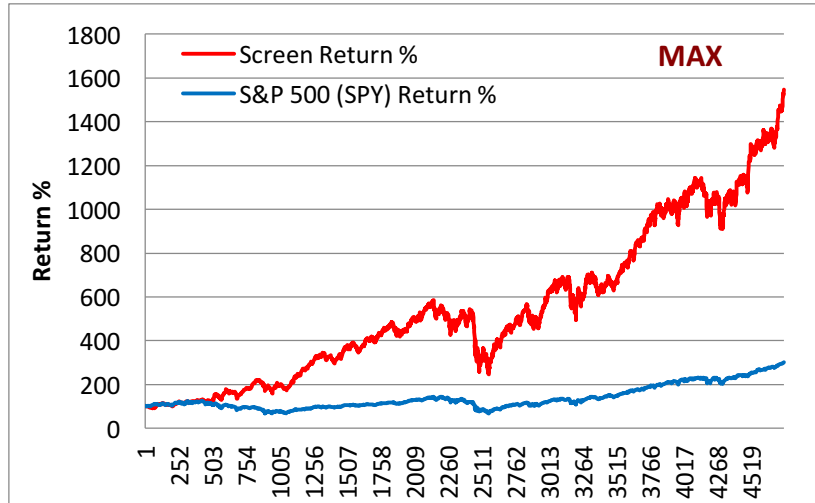
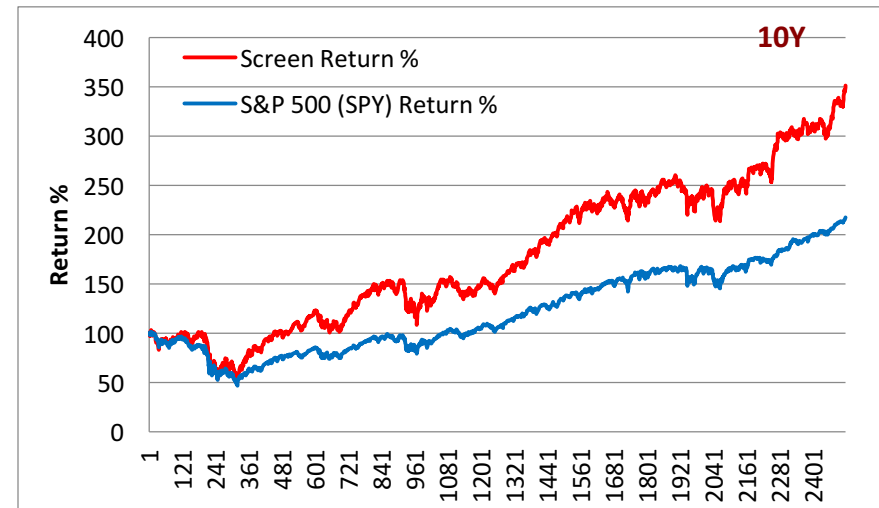
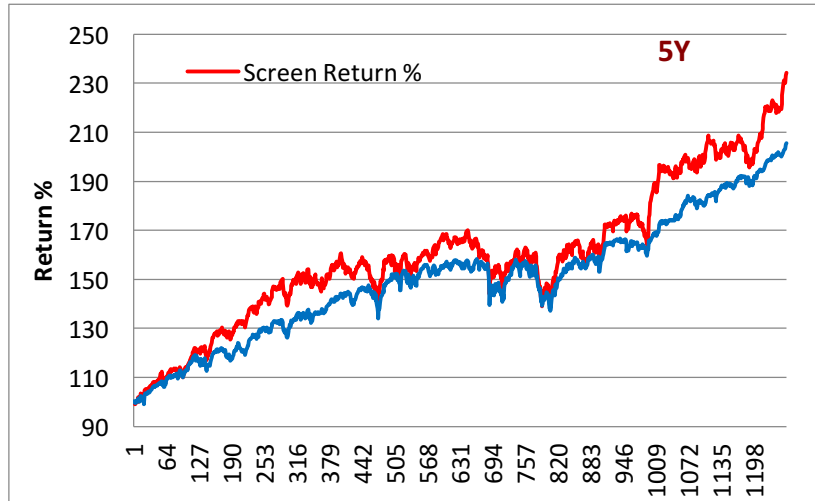


Table 2.1 Backtesting of screen with trading system in Figure 2.1

| | Return % (a.n.) | Max drawdown % | Sharpe | Sortino | StdDev % | β | α % |
|------------|--------------------|-------------------|--------|---------|-------------|---------|------------|
| S&P500-5Y | 15.50 | -13.34 | 1.37 | 1.89 | 10.58 | -- | -- |
| 5Y | 18.56 | -18.29 | 1.24 | 1.78 | 13.56 | 1.03 | 1.84 |
| S&P500-10Y | 8.11 | -54.02 | 0.54 | 0.7 | 15.76 | -- | -- |
| 10Y | 13.39 | -48.70 | 0.66 | 0.91 | 21.33 | 1.2 | 3.69 |
| S&P500-MAX | 6.02 | -55.42 | 0.33 | 0.44 | 14.84 | -- | -- |
| MAX | 15.59 | -58.24 | 0.70 | 0.97 | 21.27 | 1.21 | 9.32 |

Figure 2.7 Trading system

Prev

Re-Run Simulation

General

Buy Rules (Implicit AND)

copy to screen

| | | | |
|------------------------|---------------------------|---------|--------------------------|
| Name | Jun5_Final_Project | Size | MktCap > 200 |
| Visibility | Private | Price | close(0) > 3 |
| Category | Unclassified | Volume | AvgDailyTot(20) > 200000 |
| Starting Capital | \$1,000,000.00 | No ADRs | Universe(\$ADR) = FALSE |
| Benchmark | S&P 500 (SPY) | Buy12 | rank > 95 |
| Commission | 10.0 Flat Fee (\$) | | |
| Slippage | 0.5% of Total Amt (Fixed) | | |
| Transaction Type | Long | | |
| Use Margin | No | | |
| Management Fee | 0.0% | | |
| Price for Transactions | Next Open | | |
| Transaction Save | Yes | | |

Sell Rules (Implicit OR)

copy to screen

| | |
|-------|--------------------------------------|
| Rank | Rank < 95 // Sell low-ranking stocks |
| Price | close(0) < 2 |

Stop Loss

| | |
|----------|------|
| Strategy | None |
|----------|------|

Hedge / Market Timing DISABLED

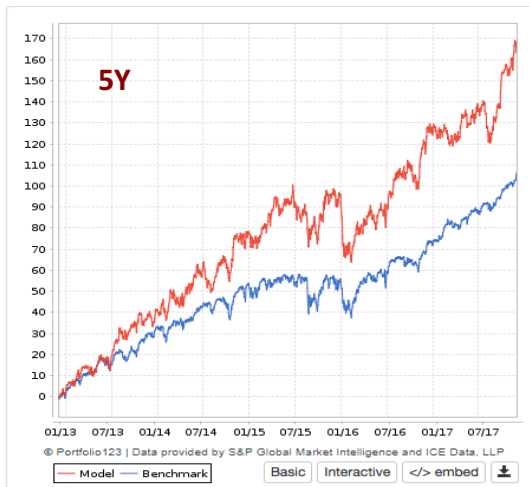
Period & Restrictions

| | |
|--------------------------|------------|
| Start Date | 01/02/1999 |
| End Date | 11/30/2017 |
| Exposure List | None |
| Restrict Buy List | |
| Restrict Sell List | |
| Load Global Restrictions | Yes |
| Allow Mergers | No |

Rebalance

Universe & Ranking

Figure 2.8 Summary of simulation



Top Holdings

| | Ticker | | Weight | Return | Shares | Value |
|----|--------|-------|--------|---------|---------|--------------|
| 1 | RTEC | 3M 1Y | 6.33% | 88.88% | 6,858.0 | \$166,649.39 |
| 2 | LIVN | 3M 1Y | 6.17% | 100.54% | 1,865.0 | \$162,590.70 |
| 3 | PLPC | 3M 1Y | 5.51% | 70.25% | 1,739.0 | \$145,189.11 |
| 4 | AZPN | 3M 1Y | 5.12% | 52.88% | 2,015.0 | \$134,843.80 |
| 5 | ATGE | 3M 1Y | 4.95% | 114.22% | 3,147.0 | \$130,443.16 |
| 6 | CNC | 3M 1Y | 4.59% | 38.04% | 1,183.0 | \$120,772.47 |
| 7 | MU | 3M 1Y | 4.47% | 38.14% | 2,780.0 | \$117,844.20 |
| 8 | NTGR | 3M 1Y | 4.36% | 52.95% | 2,230.0 | \$114,845.00 |
| 9 | VRSN | 3M 1Y | 4.11% | 149.78% | 940.0 | \$108,194.00 |
| 10 | LRCX | 3M 1Y | 3.84% | 18.76% | 526.0 | \$101,165.59 |

General Info

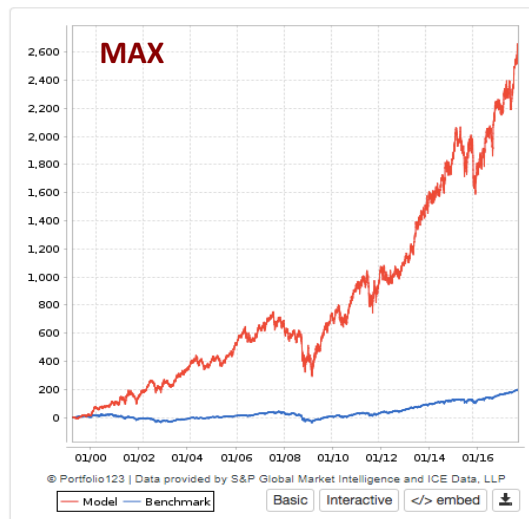
| | |
|----------------------------------|------------------------|
| Total Market Value (inc. Cash) | \$2,633,546.31 |
| Cash | \$1,057.51 |
| Number of Positions | 25 |
| Last Trades (8) | 11/20/17 |
| Period | 11/30/12 - 11/30/17 |
| Sizing Method | % Portfolio Weight |
| Last Rebalanced (Every 13 Weeks) | 11/20/17 |
| Benchmark | S&P 500 (SPY) |
| Universe | All Fundamentals - USA |
| Ranking System | June10th |

Quick Stats as of 11/30/2017

| | |
|--------------------------------|------------------|
| Total Return | 163.35% |
| Benchmark Return | 106.49% |
| Active Return | 56.86% |
| Annualized Return | 21.37% |
| Annual Turnover | 99.27% |
| Max Drawdown | -18.30% |
| Benchmark Max Drawdown | -13.02% |
| Overall Winners | (103/155) 66.45% |
| Sharpe Ratio | 1.47 |
| Correlation with S&P 500 (SPY) | 0.69 |

Recent Trades

| Date | Type | Ticker | | Shares | Price |
|----------|------|--------|-------|----------|----------|
| 11/20/17 | BUY | IDXX | 3M 1Y | 570.0 | \$154.70 |
| 11/20/17 | BUY | XOXO | 3M 1Y | 4,588.0 | \$19.24 |
| 11/20/17 | BUY | CVRR | 3M 1Y | 7,148.0 | \$12.35 |
| 11/20/17 | BUY | OSPN | 3M 1Y | 6,764.0 | \$13.05 |
| 11/20/17 | SELL | ANIK | 3M 1Y | -1,607.0 | \$53.53 |
| 11/20/17 | SELL | LMNX | 3M 1Y | -4,205.0 | \$21.50 |
| 11/20/17 | SELL | GCI | 3M 1Y | -8,403.0 | \$11.49 |
| 11/20/17 | SELL | SYNA | 3M 1Y | -2,095.0 | \$38.20 |
| 08/21/17 | BUY | PLPC | 3M 1Y | 1,739.0 | \$48.79 |
| 08/21/17 | BUY | JOUT | 3M 1Y | 1,374.0 | \$61.77 |



Top Holdings

| | Ticker | | Weight | Return | Shares | Value |
|----|---------|-------|--------|---------|-----------|----------------|
| 1 | RTEC | 3M 1Y | 6.36% | 86.13% | 71,068.0 | \$1,726,952.38 |
| 2 | ATGE | 3M 1Y | 6.20% | 146.97% | 40,613.0 | \$1,683,408.88 |
| 3 | SODA | 3M 1Y | 6.03% | 81.75% | 23,259.0 | \$1,636,968.38 |
| 4 | AZPN | 3M 1Y | 5.12% | 49.97% | 20,792.0 | \$1,391,400.62 |
| 5 | VRSN | 3M 1Y | 4.82% | 153.37% | 11,372.0 | \$1,308,917.12 |
| 6 | EXAC^18 | 3M 1Y | 4.74% | 34.87% | 30,662.0 | \$1,286,270.88 |
| 7 | CNC | 3M 1Y | 4.61% | 31.28% | 12,264.0 | \$1,252,031.75 |
| 8 | EZPW | 3M 1Y | 4.53% | 28.92% | 102,042.0 | \$1,229,606.12 |
| 9 | LRCX | 3M 1Y | 4.32% | 23.06% | 6,102.0 | \$1,173,597.62 |
| 10 | MU | 3M 1Y | 4.04% | 28.20% | 25,902.0 | \$1,097,985.75 |

General Info

| | |
|----------------------------------|------------------------|
| Total Market Value (inc. Cash) | \$27,163,862.70 |
| Cash | \$14,762.13 |
| Number of Positions | 25 |
| Last Trades (8) | 09/11/17 |
| Period | 01/02/99 - 11/30/17 |
| Sizing Method | % Portfolio Weight |
| Last Rebalanced (Every 13 Weeks) | 09/11/17 |
| Benchmark | S&P 500 (SPY) |
| Universe | All Fundamentals - USA |
| Ranking System | June10th |

Quick Stats as of 11/30/2017

| | |
|--------------------------------|------------------|
| Total Return | 2,616.39% |
| Benchmark Return | 202.80% |
| Active Return | 2,413.59% |
| Annualized Return | 19.08% |
| Annual Turnover | 134.37% |
| Max Drawdown | -53.67% |
| Benchmark Max Drawdown | -55.19% |
| Overall Winners | (379/641) 59.13% |
| Sharpe Ratio | 0.85 |
| Correlation with S&P 500 (SPY) | 0.75 |

Recent Trades

| Date | Type | Ticker | | Shares | Price |
|----------|------|---------|-------|------------|----------|
| 09/11/17 | BUY | MU | 3M 1Y | 25,902.0 | \$32.90 |
| 09/11/17 | BUY | INVA | 3M 1Y | 61,265.0 | \$13.91 |
| 09/11/17 | BUY | ACOR | 3M 1Y | 37,052.0 | \$23.00 |
| 09/11/17 | BUY | PDLI | 3M 1Y | 267,988.0 | \$3.18 |
| 09/11/17 | SELL | GORO | 3M 1Y | -250,395.0 | \$3.75 |
| 09/11/17 | SELL | CVRR | 3M 1Y | -94,427.0 | \$8.80 |
| 09/11/17 | SELL | SCMP^18 | 3M 1Y | -55,837.0 | \$12.50 |
| 09/11/17 | SELL | SGU | 3M 1Y | -88,202.0 | \$10.79 |
| 06/12/17 | BUY | EZPW | 3M 1Y | 102,042.0 | \$9.30 |
| 06/12/17 | BUY | IDXX | 3M 1Y | 5,774.0 | \$164.35 |

2.3 Simulation

After preparing the ranking system and conducting some backtest with screener, I ran simulations to inspect the performance of my trading strategy in reality. Notice that the current day is 2017-11-30. Figure 2.7 shows the trading system. Notice that I always sell the stock of rank < 95. This rule is to be consistent with the buy rule rank > 95 as well as the top one bucket performance of ranking system shown in Figure 2.3. Slippage rate is 0.5% and \$10 flat fee is imposed. Summary pages of 5Y and MAX simulation are displayed in Figure 2.8. Overall winners rates are as high as 66.45% and 59.13%, respectively.

Figure 2.9 Performance statistics

| Return (%) | 5Y | Model S&P 500 (SPY) | MAX | Model S&P 500 (SPY) |
|---------------|----|---------------------|-----|---------------------|
| Total | | 163.35 106.49 | | 2,616.39 202.80 |
| Annualized | | 21.37 15.61 | | 19.08 6.03 |
| Year To Date | | 17.37 20.25 | | 18.17 20.25 |
| Month To Date | | 3.02 3.06 | | 3.08 3.06 |
| 4 Week | | 3.67 2.88 | | 3.27 2.88 |
| 13 Week | | 13.90 7.61 | | 12.47 7.61 |
| 1 Year | | 20.67 22.68 | | 21.81 22.68 |
| 3 Year | | 50.30 35.99 | | 42.75 35.99 |

Performance by Calendar Year

5Y

| Return (%) | 2012* | 2013 | 2014 | 2015 | 2016 | 2017** |
|------------|-------|-------|-------|------|-------|--------|
| Model | 2.18 | 38.32 | 25.47 | 8.24 | 16.90 | 17.37 |
| Benchmark | 0.89 | 32.31 | 13.46 | 1.23 | 12.00 | 20.25 |
| Excess | 1.29 | 6.01 | 12.01 | 7.01 | 4.90 | -2.88 |

Performance by Calendar Year

MAX

| Return (%) | 1999* | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017** |
|------------|-------|-------|--------|--------|-------|-------|-------|-------|-------|--------|-------|-------|------|-------|-------|-------|------|-------|--------|
| Model | 78.14 | 13.06 | 43.65 | 20.09 | 33.96 | 15.30 | 15.24 | 17.82 | 4.86 | -28.01 | 49.77 | 20.76 | 2.87 | 13.48 | 38.73 | 17.68 | 7.39 | 12.91 | 18.17 |
| Benchmark | 20.39 | -9.74 | -11.76 | -21.58 | 28.18 | 10.70 | 4.83 | 15.85 | 5.15 | -36.79 | 26.35 | 15.06 | 1.89 | 15.99 | 32.31 | 13.46 | 1.23 | 12.00 | 20.25 |
| Excess | 57.75 | 22.80 | 55.41 | 41.67 | 5.78 | 4.60 | 10.42 | 1.97 | -0.29 | 8.79 | 23.42 | 5.70 | 0.97 | -2.52 | 6.42 | 4.22 | 6.15 | 0.91 | -2.07 |

(*) From 01/02/99 (**) To 11/30/17

Figure 2.9 display more performance statistics. In the return tables of both 5Y and MAX, the model was under-performed than S&P 500 in short run (i.e. year to date and month to date). Performance by calendar year also shows the capacity of the model. In addition to returns, Figure 2.10 provides risk measurement statistics. This trading strategy performs better in MAX period. For instance, the max drawdown of the model is even lower than S&P500, given the annualized alpha is as high as 13.17%. Its Sharpe ratio and Sortino ratio are also doubled or even tripled than

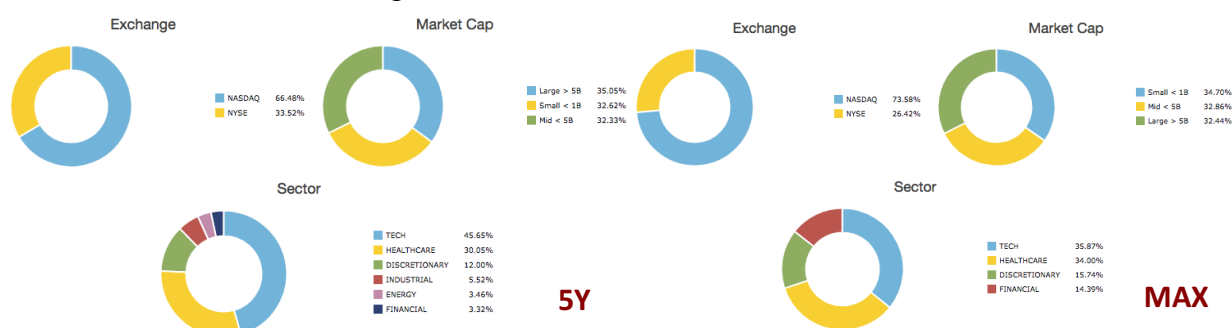
benchmark. In contrast, during 5Y period, both Sharpe ratio and Sortino ratio of the model are lower than S&P 500, indicating that the model takes more risks in return for some excess return.

Figure 2.10 Risk measurements statistics

| Since Inception 11/30/12 | | | Since Inception 01/02/99 | | |
|----------------------------|--------|---------------------|----------------------------|----------|----------------------|
| | Model | 5Y S&P 500 (SPY) | | Model | MAX S&P 500 (SPY) |
| Total Return (%) | 163.35 | 106.49 | Total Return (%) | 2,616.39 | 202.80 |
| Annualized Return (%) | 21.37 | 15.61 | Annualized Return (%) | 19.08 | 6.03 |
| Max Drawdown (%) | -18.30 | -13.02 | Max Drawdown (%) | -53.67 | -55.19 |
| Monthly Samples | 59 | 59 | Monthly Samples | 226 | 226 |
| Standard Deviation (%) | 13.52 | 9.56 | Standard Deviation (%) | 21.13 | 14.43 |
| Sharpe Ratio | 1.47 | 1.51 | Sharpe Ratio | 0.85 | 0.35 |
| Sortino Ratio | 2.06 | 2.08 | Sortino Ratio | 1.24 | 0.46 |
| Correlation with Benchmark | 0.69 | - | Correlation with Benchmark | 0.75 | - |
| R-Squared | 0.48 | - | R-Squared | 0.56 | - |
| Beta | 0.98 | - | Beta | 1.09 | - |
| Alpha (%) (annualized) | 5.94 | - | Alpha (%) (annualized) | 13.17 | - |

An inspection of portfolio allocation is shown in Figure 2.11. Market cap allocations are similar in both 5Y and MAX simulations. Specifically, the portfolios keep around equal allocations in small- mid- and large- market caps. In terms of sector allocations, both 5Y and MAX simulations focus on Tech, Healthcare, Consumer Discretionary, and Financial sectors. In addition, 5Y simulation includes small portion of investments in Industrial and Energy.

Figure 2.11 Allocation in terms of sectors



2.4 Summary

Based on the 5Y and MAX simulation results, I listed all tickers in “current holding” on 2017-11-30, as shown in Table 2.2. There are 26 stocks. I include all 26 stocks in the *raw portfolio*. In Section 4, I will discuss how to deal with the 26 stocks and come up with the *ultimate portfolio*. Before that, Section 3 is going to talk about how I collected, cleaned and constructed the key datasets I used.

Table 2.2 Current holdings on 2017-11-30

These stocks are the current holdings on 2017-11-30, according to 5Y and MAX simulation results. Green panel consists of 15 tickers appearing on both 5Y and MAX lists. Orange panel consists of 6 tickers appearing only on 5Y list. Blue panel consists of 5 tickers appearing only on MAX list.

| Ticker | Weight | Return | Rank | Days Held | Sector* |
|--------|--------|---------|------|-----------|-------------|
| ACOR | 3.00% | -7.35% | 99 | 101 | Health Care |
| ATGE | 4.95% | 114.22% | 96.7 | 556 | CD |
| AZPN | 5.12% | 52.88% | 98.2 | 919 | IT |
| CNC | 4.59% | 38.04% | 98.2 | 192 | Health Care |
| CRUS | 2.93% | -4.83% | 98.2 | 374 | IT |
| IDXX | 3.39% | 0.59% | 99.6 | 10 | Health Care |
| INVA | 3.47% | 6.99% | 98.3 | 101 | Health Care |
| LCI | 3.77% | 22.40% | 99.1 | 374 | Health Care |
| LRCX | 3.84% | 18.76% | 99.3 | 101 | IT |
| MU | 4.47% | 38.14% | 99.8 | 101 | IT |
| NTGR | 4.36% | 52.95% | 98.9 | 1283 | IT |
| RTEC | 6.33% | 88.88% | 98.1 | 919 | IT |
| SIRI | 3.23% | 39.59% | 97.5 | 556 | CD |
| UTHR | 2.46% | -14.65% | 99.6 | 738 | Health Care |
| VRSN | 4.11% | 149.78% | 95.3 | 1739 | IT |
| JOUT | 3.82% | 17.80% | 98.8 | 101 | CD |
| NVMI | 3.82% | 17.94% | 99.6 | 101 | IT |
| PLPC | 5.51% | 70.25% | 99 | 101 | Industrials |
| TARO | 3.20% | 4.02% | 96.4 | 374 | Health Care |
| TER | 3.84% | 18.42% | 97.6 | 101 | IT |
| XOXO | 3.37% | -0.04% | 99.7 | 10 | IT |
| EZPW | 4.53% | 28.92% | 99 | 171 | Financials |
| GBL | 2.91% | -12.26% | 97.3 | 353 | Financials |
| KLIC | 4.03% | 14.81% | 98.5 | 171 | IT |
| LMNX | 3.57% | 1.69% | 98.3 | 171 | Health Care |
| PDLI | 2.87% | -8.95% | 99.9 | 80 | Health Care |

* CD – Consumer Discretionary; IT – Information Technology

3. Numbers and Text Dataset

3.1 Data Collection and Data Cleaning

I collected both *numbers* and *text* dataset from Jan 1st, 2007 to Apr 30th, 2018. The *numbers* dataset contains stock prices, Fama-French 3 factors and CPI. The *text* is Reuters news

archive which was scraped with Python codes. I ran the code on midway terminals from May 17th to June 5th, 2018. Finally, 3,279,343 of news has been collected. During the process of news scraping, some news were missing due to webpage error or non-existing news links.

In terms of *text* dataset, after scraping all news, I firstly tokenized each headline and removed the words appearing on the stop list. Then I removed the punctuations. Afterwards, the rest of tokens were lemmatized. These three steps reduced the dimensionality of the raw corpus from 30,729,641 to 24,857,489. Finally, I transformed all words into lower case.

3.2 Dataset Construction

There were two major data frames being constructed.

I. *Prices.csv*

This data frame includes all dependent variables, i.e. both excess returns and *IAERs* of all 26 stocks filtered in Section 2. *IAER* is formulated as follows:

$$IAER = (Price_1/CPI_1 - Price_0/CPI_0) / (Price_0/CPI_0)$$

IAER is calculated on a daily-basis. However, only monthly CPI data is available. I used weighted average to transform monthly CPI into daily CPI. Briefly speaking, I assigned available monthly CPI to the 15th of each month. Any other day's CPI is calculated by weighted-averaging the previous month's 15th CPI and next month's 15th CPI values. The weights are calculated based on the number of days between the day and the previous month's 15th, and the number of days between the day and the next month's 15th. Table 3.1 displays an example of *Prices.csv*.

Table 3.1 *Price.csv*

| Prices_CSV[:5] | | | | | | | |
|----------------|-----------|-----------|----------|-----------|-----------|-----------|----------|
| ACOR_IAER | ATGE_IAER | AZPN_IAER | CNC_IAER | CRUS_IAER | IDXX_IAER | INVA_IAER | LCI_IAER |
| -1.857 | -1.058 | 3.823 | 0.580 | 2.670 | 1.304 | 0.542 | 0.284 |
| 1.350 | 0.908 | -3.880 | 0.618 | 2.228 | 2.726 | 0.540 | -0.090 |
| 0.699 | -0.691 | -0.272 | 0.414 | -0.272 | 0.654 | 0.204 | -1.226 |
| 1.955 | 0.283 | 0.627 | 0.088 | -0.032 | -0.154 | 0.946 | -2.760 |
| 1.372 | -0.052 | -1.383 | -2.022 | 2.752 | -0.887 | -0.668 | -4.057 |

II. Factors.csv

This data frame includes all independent variables, i.e. both financial factors and sentiment factors. There are three financial factors, namely SMB, HML, and market excess returns (i.e. Fama-French 3 factors). There are 14 sentiment factors.

Sentiment factors were made based on four dictionaries, namely textBlob, AFINN, BING and NRC. Specifically, textBlob is a python package returns a polarity value to any input bag of words. The polarity ranges from -1.0 to 1.0. Negative polarity value represents negative sentiment. Similarly, AFINN assigns sentiment score ranging from -5.0 to 5.0. BING characterizes each word into “negative” or “positive”. NRC associates each word with one of ten emotions, including anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise and trust.

To extract the sentiment, I zoomed in each day and counted the number of words under different categories assigned by the three dictionaries. For example, on 2007-01-04, there are 142 *anger* words, 10 *disgust* words, 61 *fear* words, 37 *joy* words, 120 *negative* words, 215 *positive* words, 5 *sadness* words appearing in that day’s news headlines. Afterwards, I normalized the number of words by dividing them by the total number of words. More details of creating sentiment factors is in Appendix – construct sentiment factors.

4. Model Training, Testing and Predicting

With the *training* dataset from 2007-01-03 to 2015-09-09, I trained such models as linear regression, LASSO, partial least squares (PLS), CART Tree, and Random Forest on each selected 26 stocks. The independent variables are lag-time series of factors. On the one hand, considering that I need to predict the stock performances in *future*, using the data from current day is unrealistic, given the fact that collecting and synthesizing data requires time. Otherwise, the trading actions may not be responsive due to waiting for the up-to-date data. On the other hand, time series data usually displays some momentum in continuing (or changing slightly from) previous days status. So, instead of using the current day SMB, HML, all sentiment factors data, I generated 2-, 3-, 4-, 5-, 10-, 15-, 20-, 30-lag time series. Meanwhile, I also transferred the dependent variables, i.e. *IAER*, to binary “1” or “-1”. I just focused on whether the *IAER* was positive or negative. Hence, the predictions are whether the *IAER* of a certain stock will be positive or negative in *future*.

To test the models, I applied each of them on *testing* dataset from 2015-09-10 to 2017-11-30. Table 4.1 summarizes the correction rates of each model. The correction rate is incidence of models making the correct predictions as the actual testing data shows. On each row of Table 4,1, I highlighted two models with the highest correction rates. However, I did not consider the stickers whose corresponding correction rate was below 0.5. Therefore, stocks ATGE, CNC, IDXX, MU and KLIC were no longer considered.

Table 4.1 Correction rate of models with *testing* dataset
Best two models are highlighted.

| | Linear Regression | LASSO | PLS | CART Tree | Random Forest |
|------|----------------------|--------|--------|-----------|---------------|
| ACOR | 0.5229 | 0.4844 | 0.5358 | 0.4844 | 0.4862 |
| ATGE | 0.4844 | 0.4954 | 0.4862 | 0.4789 | 0.4826 |
| AZPN | 0.5303 | 0.4789 | 0.5229 | 0.4606 | 0.5211 |
| CNC | 0.4954 | 0.4917 | 0.4936 | 0.4917 | 0.4771 |
| CRUS | 0.5064 | 0.5138 | 0.5064 | 0.5138 | 0.4642 |
| IDXX | 0.4881 | 0.4220 | 0.4991 | 0.4459 | 0.4826 |
| INVA | 0.5284 | 0.4936 | 0.5229 | 0.4936 | 0.5138 |
| LCI | 0.5083 | 0.5376 | 0.5083 | 0.5413 | 0.5376 |
| LRCX | 0.5413 | 0.4459 | 0.5450 | 0.4459 | 0.5156 |
| MU | 0.4789 | 0.4917 | 0.4679 | 0.4936 | 0.4789 |
| NTGR | 0.5064 | 0.5064 | 0.5028 | 0.5064 | 0.4826 |
| RTEC | 0.5101 | 0.4807 | 0.5028 | 0.4807 | 0.4917 |
| SIRI | 0.5083 | 0.4752 | 0.5028 | 0.4752 | 0.5211 |
| UTHR | 0.5046 | 0.5303 | 0.5119 | 0.5303 | 0.5119 |
| VRSN | 0.5486 | 0.4679 | 0.5468 | 0.4679 | 0.4899 |
| JOUT | 0.4936 | 0.4917 | 0.4936 | 0.4917 | 0.5064 |
| NVMI | 0.4606 | 0.5028 | 0.4697 | 0.5028 | 0.5083 |
| PLPC | 0.4661 | 0.5284 | 0.4679 | 0.5284 | 0.4972 |
| TARO | 0.5321 | 0.5321 | 0.5321 | 0.5376 | 0.5229 |
| TER | 0.5064 | 0.4330 | 0.5046 | 0.4330 | 0.4991 |
| XOXO | 0.5211 | 0.5119 | 0.5138 | 0.5119 | 0.4569 |
| EZPW | 0.4807 | 0.5083 | 0.4789 | 0.5083 | 0.4936 |
| GBL | 0.5064 | 0.5468 | 0.5156 | 0.5468 | 0.4569 |
| KLIC | 0.4826 | 0.4972 | 0.4881 | 0.4972 | 0.5028 |
| LMNX | 0.4752 | 0.5266 | 0.4716 | 0.5266 | 0.4789 |
| PDLI | 0.5596 | 0.5009 | 0.5761 | 0.5046 | 0.4826 |

With the *predicting* dataset from 2017-12-01 to 2018-04-30, I predicted each stock's *IAER* with the corresponding best-performed models as highlighted in Table 4.1. There are two models in most cases, the predicting results were only considered when two models gave the same prediction (i.e. achieving agreement). Table 4.2 displays the predicting results. Based on the results, four stocks (i.e. AZPN, LRCX, SIRI, TER) firstly secured their positions in the *ultimate portfolio*, given their higher probability of generating more positive *IAER* in *future*. Analogously, such stocks as CRUS, LCI, NTGR, UTHR, NVMI, PLPC, EZPW, GBL, LMNX lost their chance to enter the *ultimate portfolio*. Therefore, in next section, I will consider the four secured stocks AZPN, LRCX, SIRI, TER, and ten candidate stocks JOUT, ACOR, INVA, RTEC, VRSN, JOUT, TARO, XOXO, KLIC, and PDLI.

Table 4.2 Predicting performance of stocks

| | Models | Agreement Rate | Predictions | |
|------|---------------------|-------------------|-------------|------------|
| | | | # (-) IAER | # (+) IAER |
| ACOR | Linear, PLS | 0.9857 | 35 | 34 |
| AZPN | Linear, PLS | 0.9857 | 22 | 47 |
| CRUS | LASSO, Tree | 1 | 70 | 0 |
| INVA | Linear, PLS | 0.929 | 31 | 34 |
| LCI | LASSO, Tree, RF | 1 | 70 | 0 |
| LRCX | Linear, PLS | 0.9857 | 17 | 52 |
| NTGR | Linear, LASSO, Tree | 1 | 70 | 0 |
| RTEC | Linear, PLS | 0.9571 | 37 | 30 |
| SIRI | Linear, PLS | 0.9286 | 26 | 39 |
| UTHR | LASSO, Tree | 1 | 70 | 0 |
| VRSN | Linear, PLS | 0.7857 | 32 | 23 |
| JOUT | RF | -- | 46 | 24 |
| NVMI | LASSO, Tree | 1 | 70 | 0 |
| PLPC | LASSO, Tree | 1 | 70 | 0 |
| TARO | Tree | -- | 39 | 31 |
| TER | Linear, PLS | 0.9857 | 20 | 49 |
| XOXO | Linear | -- | 39 | 31 |
| EZPW | LASSO, Tree | 1 | 70 | 0 |
| GBL | LASSO, Tree | 1 | 70 | 0 |
| KLIC | RF | -- | 39 | 31 |
| LMNX | LASSO, Tree | 1 | 70 | 0 |
| PDLI | Linear, PLS | 1 | 36 | 34 |

5. Portfolio Creation

5.1 The *Ultimate Portfolio*

In my opinion, portfolio is not a simple combination of individual stocks, but also including all the interactions among stocks, and stocks with outside environment changes. Therefore, I did not stop after sorting out the four stocks AZPN, LRCX, SIRI, TER and putting them into the *ultimate portfolio*. I continued to adjust the portfolio by minimizing the correlations among stocks. According to risk diversification, a portfolio with various stocks negatively correlated with one another tend to expose to less risks. The formula below models the relationship between the portfolio and its stocks. $2 \sum_{i=1}^n \sum_{j \neq i} w_i w_j \text{Cov}(X_i, X_j)$ summarize the interactions of individual stocks. Negative correlations among stocks contribute to reducing portfolio's variance, a measure of risks taking by the portfolio. Moreover, as the number of stocks increasing, $2 \sum_{i=1}^n \sum_{j \neq i} w_i w_j \text{Cov}(X_i, X_j)$ dominates the value $\text{Var}(P)$. Hence, such rationale of diversifying risk motivates me to refine my portfolio by minimizing the portfolio correlation among stocks.

$$\text{Var}(P) = \text{Var}(w_1 X_1 + \dots + w_n X_n) = \sum_{i=1}^n w_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j \neq i} w_i w_j \text{Cov}(X_i, X_j)$$

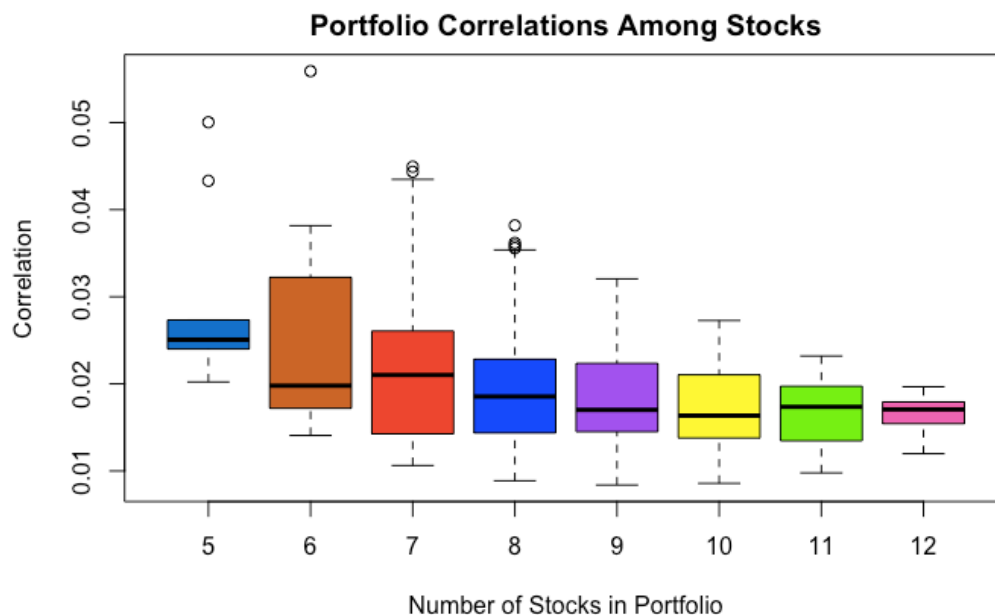
With the four stocks already being in the *ultimate portfolio* and ten stocks as candidates, I tried all the combinations of stocks to find out the one with the most negative (or least positive) portfolio correlation among stocks. Table 5.1 displays an example of portfolio correlations.

Table 5.1 Correlation of stocks in portfolios
Portfolio = {AZPN, LRCX, SIRI, TER, + n}

| + one more stock | Correlations $\sum_{i=1}^n \sum_{j \neq i} w_i w_j \text{Cov}(X_i, X_j)$ |
|-----------------------------------|---|
| RTEC_IAER | 0.043312 |
| VRSN_IAER | 0.027336 |
| XOXO_IAER | 0.024771 |
| [ACOR_IAER, TARO_IAER] | 0.014599 |
| [ACOR_IAER, XOXO_IAER] | 0.018240 |
| [ACOR_IAER, XOXO_IAER] | 0.018240 |
| [INVA_IAER, KLIC_IAER, PDLI_IAER] | 0.026850 |
| [RTEC_IAER, VRSN_IAER, JOUT_IAER] | 0.021846 |
| [RTEC_IAER, VRSN_IAER, TARO_IAER] | 0.018910 |
| [RTEC_IAER, VRSN_IAER, XOXO_IAER] | 0.025590 |

Figure 5.1 displays the correlation distributions by the number of stocks in portfolios. It substantiates that involving more stocks tends to reduce the portfolio correlation. However, among all the 512 portfolio candidates, none of portfolio has correlation of stocks being negative. Instead, I found the portfolio with the smallest correlation of stocks. This portfolio contains nine stocks AZPN, LRCX, SIRI, TER, ACOR, VRSN, JOUT, TARO, and XOXO with the smallest correlation $\sum_{i=1}^n \sum_{j \neq i} w_i w_j \text{Cov}(X_i, X_j) = 0.00839$. Therefore, I define it as the *ultimate portfolio*. Recall Table 2.3, ACOR and TARO are from health care, AZPN, LRCX, TER, VRSN and XOXO are from IT, SIRI, JOUT are from consumer discretionary. Moreover, AZPN has ranking 98.2, LRCX 99.3, SIRI 97.5, TER 97.6, ACOR 99, VRSN 95.3, JOUT 98.8, TARO 96.4, XOXO 99.7.

Figure 5.1: Portfolio correlations among stocks vs. portfolio size



Furthermore, if I did not reserve positions for stocks AZPN, LRCX, SIRI, and TER, but instead accepted any combination of stocks, portfolio of RTEC, SIRI, VRSN and TARO has stock correlation -0.01013; and portfolio of RTEC, SIRI, VRSN, JOUT, and TARO has stock correlation -0.006768.

5.2 Predicting the Performance of the *Ultimate Portfolio*

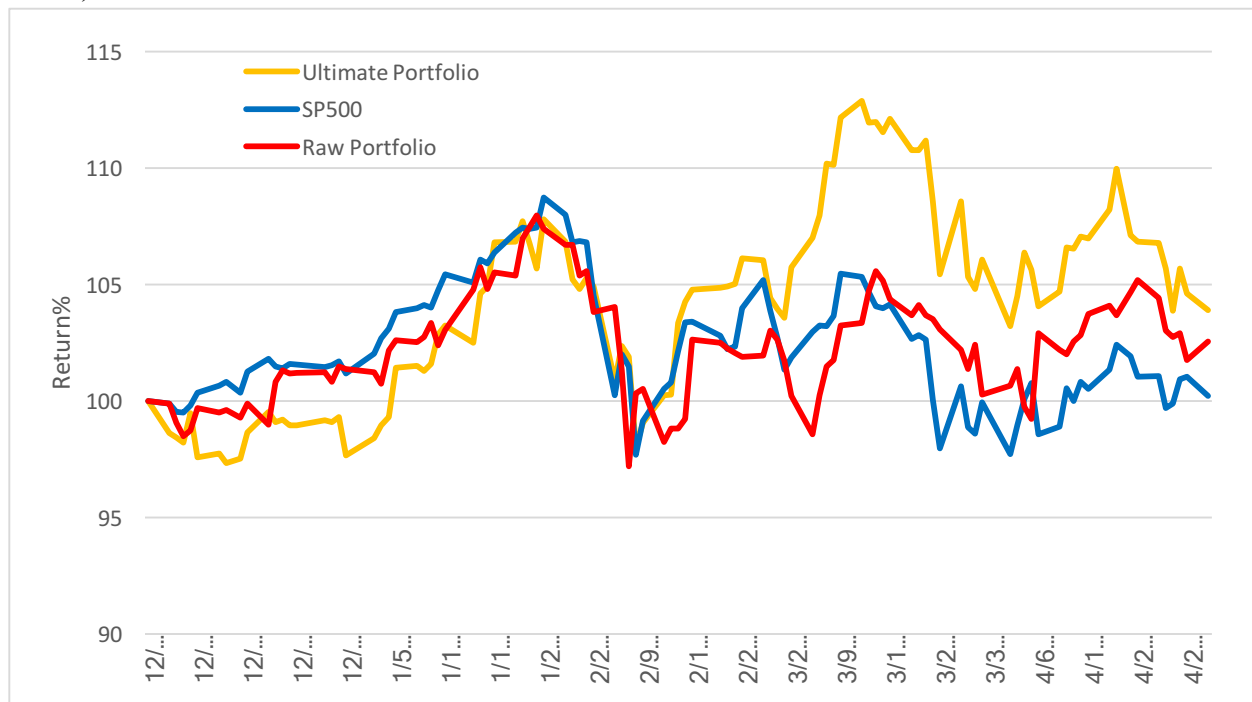
Recall that *today* is 2017-11-30. Based on my analysis on dataset from 2007-01-04 to 2017-11-30, I created the *raw portfolio* containing twenty-six stocks. Then with various statistical models and sentiment analysis, I predicted the stocks' performance in *future* 2017-12-01 to 2018-

04-30, and finally generated the *ultimate portfolio* containing nine stocks, namely AZPN, LRCX, SIRI, TER, ACOR, VRSN, JOUT, TARO, and XOXO. Hopefully, the ultimate portfolio will perform well in *future*.

Now, taking the time machine, I obtained the actual stock/index adjusted close prices in *future* from 2017-12-01 to 2018-04-30. To backtest and compare the performance of *raw portfolio*, *ultimate portfolio* and benchmark SP500, I used the actual prices to calculate portfolio returns. Figure 5.2 displays the results. Before Feb, 2018, three portfolios moved close to each other, while S&P 500 surprisingly beating the other two portfolios. After Feb, 2018, the *ultimate portfolio* greatly outperformed the other two counterparties. The raw portfolio, although underperforming than the *ultimate portfolio*, started to beat the market S&P500 since mid of Mar, 2018. Therefore, I succeeded in positively managing portfolios, finding the *raw portfolio* beating the market, and refining it to an even better performed *ultimate portfolio* in *future*.

Figure 5.2 *Ultimate portfolio, raw portfolio and S&P500*
Performance during 2017-12-01 ~ 2018-04-30

Raw portfolio contains 26 stocks selected through Portfolio123 rnakign and screening system. Ultimate portfolio is weighted-average portfolio containing stocks AZPN, LRCX, SIRI, TER, ACOR, VRSN, JOUT, TARO, and XOXO



6. Conclusion

This paper aims at creating portfolio outperforming the market S&P 500. The rationale of my strategies are (1) buying wonderful business at bargaining price with high momentum and strong industrial leadership (2) integrating market sentiment extracted from news *text* into quantitative analysis (3) further picking up stocks with greater probability to generate more positive inflation-adjusted excess returns (4) portfolio is not a simple summation of individual stocks, but an organic system with stocks interactions and their responses to outside environment changes. Taking portfolio correlation among stocks into account improves the performance of the portfolio.

With these four major rationale, I created the ranking and screening system on Portfolio123. By adjusting the components, parameters and weights, I found the trading system yielding robust excess returns under different time frames, namely 5Y, 10Y, and MAX. with this system, I made the *raw portfolio*. Then I applied linear regression, LASSO, partial least squares (PLS), CART Tree and random forest to predict the stock performances in *future* with financial and sentiment factors. Those with negative predictions were discarded. With the remaining twenty-one stocks, I inspected their interactions by filtering out the combination with the least portfolio correlation among stocks. This combination is the *ultimate portfolio* consisting of nine stocks. These nine stocks come from three sectors, namely IT, health care and consumer discretionary. The actual performance of *ultimate portfolio* is satisfying. It not only outperformed the raw S&P 500, but also beat the *raw portfolio*.

In terms of robustness, there are three main actions I kept taking to guarantee the robustness of my results. Firstly, the trading system on Portfolio 123 functions well in various time frames, including 5Y, 10Y, MAX. Secondly, I used four different dictionaries while measuring the market sentiment, namely TextBlob, AFINN, BING and NRC. Thirdly, I applied various models such as linear regression, LASSO, PLS, CART Tree, and random forest to train and predict the *IAERs*.

Last but not least, there are still open problems on how to deal with cyclical variations, seasonality, irregular movements over years, especially the time frame of dataset gets longer. These potential problems may cause structural changes in macro indicators, market sentiment and financial factors. They are yet fully considered in this paper.

Appendix – construct sentiment factors

Dictionary 1: *AFINN*

For each day, I went through all the tokens, and assigned tokens positive or negative values according to AFINN dictionary, given that the tokens were on the AFINN list. Those words off AFINN were recorded as missing words. Then I sum up negative values of all negative words to get the negative score for that day. I got the positive score for that day through a similar way. Then I calculated the compounded score by summing up negative and positive scores. Finally, I normalized the both negative and positive scores by dividing the sum of negative word number and positive word number. Table A1 shows an example of my AFINN analysis.

Table A1: An example of *AFINN* structure

| | Date | Positive_score | Negative_score | Positive_words | Negative_words | Missing_words |
|---|------------|----------------|----------------|--|--|---|
| 0 | 2007-01-04 | 270 | -304 | straight share best help cool top peace growth... | crush warn alone cut cut crisis no regret pay ... | leafs score nine goal bruises shareholder vodaf... |
| 1 | 2007-01-05 | 333 | -368 | warm successful commits boosting romance peace... | suicide cut cut warn poor kill resign dead cut... | singh move one ahead wet windy kapalua some do... |
| 2 | 2007-01-08 | 478 | -557 | fresh boost help growth hope ease big big posi... | unhappy lonely lonely pressure risk collide in... | press digest washington post business jan iraq... |
| 3 | 2007-01-09 | 302 | -665 | resolve top interest top fame vitamin solid st... | debt infringement drag flu miss disaster disas... | update file patent suit china give hk pandas m... |
| 4 | 2007-01-10 | 425 | -647 | justice share boost chance big awards expands ... | murder poor death disaster disaster drop criti... | press digest financial times jan india pantalo... |
| 5 | 2007-01-11 | 397 | -619 | share comedy peacefully marvel boost success f... | battle injury worry dead miss weakness anti no... | japan topix rise pct tech bank factbox players... |

Dictionary 2: *BING*

For each day, I went through all the tokens, and classified tokens as *positive* or *negative* according to BING dictionary, given that the tokens were on the BING list. Those words off BING list were recorded as missing words. I counted the number of negative and positive words. I calculated the compounded score by subtracting the number of negative words from the number of positive words. Then, I normalized the compounded score by dividing the sum of the negative words number and positive words number. Table A2 gives an example of my BING analysis.

Dictionary 3: *NRC*

For each day, I went through all the tokens, and classified tokens as *anger*, *anticipation*, *disgust*, *fear*, *joy*, *negative*, *positive*, *sadness*, *surprise* or *trust* according to NRC dictionary, given

that the tokens were on the NRC list. Those words off NRC list were recorded as missing words. Then I calculated the compounded numbers of words in each category. I standardized the compounded numbers by dividing total number of words (excluding missing words). Table A3 gives an example of my NRC analysis.

Table A2: An example of *BING* structure

| | Date | Positive_num | Negative_num | Missing_num | Positive_words | Negative_words | Missing_words |
|---|------------|--------------|--------------|-------------|---|---|---|
| 0 | 2007-01-04 | 141 | 222 | 4193 | best cool blossom top peace gain best portable... | crush fall stigma crisis regret slow debt fall... | leafs score nine straight goal bruises sharehol... |
| 1 | 2007-01-05 | 166 | 253 | 3461 | warm successful intelligence lead renewed peac... | mistakenly tentative sue suicide poor kill rad... | singh move one ahead wet windy kapalui some do... |
| 2 | 2007-01-08 | 207 | 340 | 5322 | fresh boost wonder steady ease positive positi... | resistance resistance unhappy bleeds lonely fl... | press digest washington post business jan iraq... |
| 3 | 2007-01-09 | 169 | 425 | 5185 | lead top top idol fame lighter solid strong to... | debt infringement fall cheap drag miss disaste... | update resolve file patent suit china give hk ... |
| 4 | 2007-01-10 | 198 | 381 | 5787 | boost awards boost tranquil best top guarantee... | murder poor death fall grim disaster disaster ... | press digest financial times jan india pantalo... |

Table A3: An example of *NRC* structure

| | Date | Anger_num | Anticipate_num | Disgust_num | Fear_num | Joy_num | Negative_num | Positive_num | Sadness_num | Surprise_num | Trust_num |
|---|------------|-----------|----------------|-------------|----------|---------|--------------|--------------|-------------|--------------|-----------|
| 0 | 2007-01-04 | 142 | 0 | 10 | 61 | 37 | 120 | 215 | 5 | 10 | 56 |
| 1 | 2007-01-05 | 167 | 0 | 18 | 74 | 34 | 131 | 241 | 0 | 16 | 63 |
| 2 | 2007-01-08 | 226 | 0 | 25 | 137 | 41 | 120 | 325 | 5 | 24 | 58 |
| 3 | 2007-01-09 | 247 | 0 | 34 | 139 | 38 | 179 | 241 | 3 | 16 | 104 |
| 4 | 2007-01-10 | 282 | 0 | 27 | 133 | 53 | 156 | 338 | 3 | 15 | 124 |
| 5 | 2007-01-11 | 259 | 0 | 27 | 154 | 54 | 173 | 331 | 5 | 13 | 135 |

Finally, I combined all the sentiment scores in the *Sentiment_Factor.csv* as shown in Table A4. Last but not list, the time frames of dataset *Stock_and_Index.csv*, *Word_Frequency.csv*, and *Sentiment_Factor.csv* matched with one another through variable “Date”.

Table A4: An example of *Sentiment_Factor.csv* structure

| | TextBlob | AFINN_Positive | AFINN_Negative | BING_Positive | BING_Negative | NRC_Anger | NRC_Disgust | NRC_Fear | NRC_Joy | NRC_Negative | NRC_Positive |
|---|----------|----------------|----------------|---------------|---------------|-----------|-------------|----------|---------|--------------|--------------|
| 0 | 0.054109 | 0.828 | -0.933 | 0.388 | 0.612 | 0.216 | 0.015 | 0.093 | 0.056 | 0.183 | 0.328 |
| 1 | 0.077918 | 0.854 | -0.944 | 0.396 | 0.604 | 0.224 | 0.024 | 0.099 | 0.046 | 0.176 | 0.324 |
| 2 | 0.048532 | 0.872 | -1.016 | 0.378 | 0.622 | 0.235 | 0.026 | 0.143 | 0.043 | 0.125 | 0.338 |
| 3 | 0.039919 | 0.812 | -1.788 | 0.285 | 0.715 | 0.247 | 0.034 | 0.139 | 0.038 | 0.179 | 0.241 |
| 4 | 0.045828 | 0.878 | -1.337 | 0.342 | 0.658 | 0.249 | 0.024 | 0.118 | 0.047 | 0.138 | 0.299 |
| 5 | 0.045266 | 0.820 | -1.279 | 0.365 | 0.635 | 0.225 | 0.023 | 0.134 | 0.047 | 0.150 | 0.288 |