Improved Anomalies Strategy with Sentiment Data
MS&E 448 Final Report

Yu Gu, Ruoyu Han, Yuzhu Zhang

June, 2020

Abstract

The project explores the impact of well-known market anomalies and constructs a set of useful factors to predict the market performance. We divided the signals into 7 classes: market factors, momentum factors, value factors, growth factors, profitability factors, liquidity factors and sentiment factors. Based on these traditional anomalies, we selected 38 classical factors and 29 sentiment factors for better anomalies prediction and have combined these signals to build robust portfolios. We first analyzed individual signal's performance and selected the top factors based on the Sharpe ratio of their single factor's portfolios and re-evaluate the factors each year. Then we built predictive models with these factors and our final portfolio achieves an annualized Sharpe ratio of 0.8909 using constructed classical factors and 0.9903 using sentiment factors together with classical factors.

1 Introduction

Well-know market anomalies happen when one particular security or group of securities perform contrary to the notion of an efficient market, where security prices are said to reflect all available information at any point in time. Typical anomalies include small firms, and low volatility or high book-to-price stocks tend to outperform. And stocks underperformed in the fourth quarter last year tend to outperform in the coming January, which is known as the January effect. Therefore, many investors make use of these market anomalies to form popular trading strategies. In an influential paper, Fama and French [1993] developed a 3-factor model, consisting of the market excess return factor, the size premium factor(small-minus-big, SMB), and the value premium factor(high-minus-low, HML), summarizes the cross-section of average stock returns as of the mid-1990s. After that, many researchers also summarized these anomalies in the stock market into different aspects such as size, value, volatility, quality, and momentum, and has shown their effectiveness.(Bouchaud et al. [2016], Ciliberti et al. [2019], Beveratos et al. [2017], Blanc et al. [2014])

During our preliminary analysis, we constructed portfolios from 2009 to 2018 based on the following factors mentioned in these papers(Fama and French [1993], Bouchaud et al. [2016], Ciliberti et al. [2019], Beveratos et al. [2017], Blanc et al. [2014]).

- Size factors: SMB(small capitalization minus big capitalization), CMH(cold minus hot, average daily volume)
- Quality factors: ROA(high return-over-assets minus low return-over-assets) , OCF(high net operating cash flow minus low net operating cash flow)
- Volatility factor: LowVol(low volatility minus high volatility)
- Momentum factor: UMD(up minus down momentum)
- Value factor: HML (high book-to-price minus low book-to-price)

However, these factors did not perform quite well and we concluded it could be the following reasons. Firstly, we computed daily factors in which the changes between a longer period may not be captures. Besides, traditional factors like CMH and HML suggested in Fama and French [1993] are outdated for now. Moreover, the factors are time-dependent and may only be effective during a specific short period,

therefore we should not apply the same factors for all years. Therefore in our further steps, we computed monthly factors, re-evaluated, and updated the efficient ones for every single year. To capture more effective factors, we computed 38 classical signals in six categories: market, momentum, value, growth, profitability, liquidity.

Besides, with the help of natural language processing techniques and the emerging platforms for investors to post their comments(such as Twitter and Stocktwits), sentiment data has also been proved to help predict the trend in the stock market(Dhaoui and Bensalah [2017], Chung et al. [2012], Kim and Kim [2014], Baker and Wurgler [2006], Berger and Turtle [2012], Blanc et al. [2014]), since it can reflect the mood of investors and some of the anomalies are indeed psychologically driven. Therefore, we also added 29 sentiment factors, including 19 direct and 10 indirect ones, to improve our market prediction. To test the predicted power of our selected factors, we built seven prediction models and it was shown in the results section that our approach is quite effective.

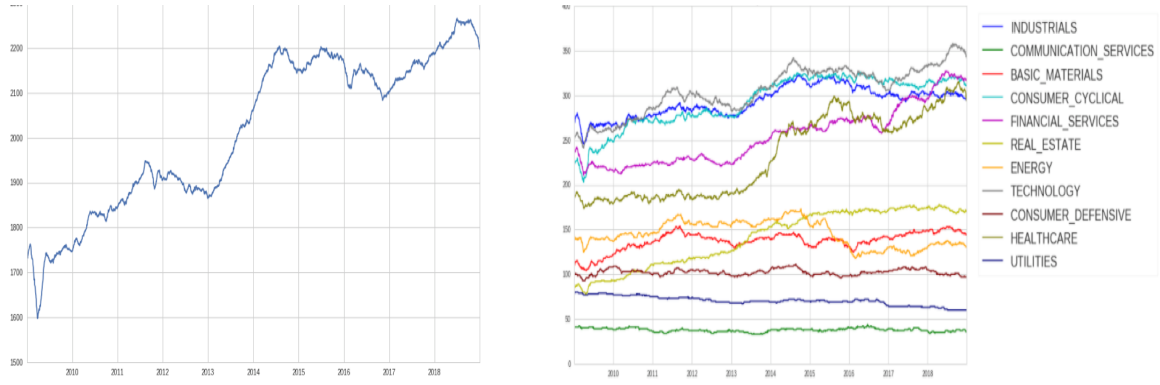
The paper is then divided into the following sections: in Section 2 we introduce the data universe; Section 3 discusses the selection of our factors; Section 4 introduces our approach to analyze the factors; Section 5 includes our experiments and the results of our models are shown in Section 6; Section 7 gives the final conclusion and possible future work of our project.

2 Data Universe

For the data universe, we selected QTradableStocksUS from Quantopian because it has the following characteristics:

- It is a reliable resource with no survivor bias
- It provides a set of liquid, easy-to-trade stocks while excluding assets that have more difficult risk profiles like ADRs and ETFs
- No explicit size limit, and generally has between 1600-2100 members each day.

For our project, we want data to be as much as possible and therefore we selected a 10 years period from 2009-1-1 to 2018-12-31 to inspect. A visualization of the universe is shown in Figure 1.



(a) Number of securities

(b) Number of securities per sector

Figure 1: QTradableStocksUS universe from 2009 to 2018

3 Factors Selection

3.1 Classical Factors

We created 38 classical monthly factors. They can be broadly categorised into market, momentum, value, growth, profitability and liquidity metrics.

Size, momentum and value anomalies are well-explored since Carhart [1997] published four factor model based on the Fama and French three factor model. However, instead of using SMB and HML, we extracted the value of each stock as its factor. In addition to those factors, we created more categories. For market factors, we captured characters of dynamic stock price, such as volatility and turnover. As for the growth factor, we computed the growth rate of important indexes of one company. In terms of profitability and liquidity factor, we extracted the most common financial ratios in quarterly report that can reflect the profitability and liquidity of one company. Both of them are vital for the enterprises.

The detailed information about factors are in tables below and Appendix.

Market	Momentum	Value	Growth	Profitability	Liquidity
Size	Mom6	BP	Asset Growth	ROA	Current_Ratio
Beta,Betasq	Mom12	EP	EPS_G	ROE	Quick_Ratio
Vol	Mom36	CFP	DPS_G	Asset_Turnover	CF_Sales
Skew	Momchg	SP	BPS_G	FCF_Yield	
Turn, std_turn	Lagretn	PEG	NI_G		
Volume, Volume_std		EV_Ebitda	OI_G		
Maxetn			Sales_G		
Sharechg			Revenue_G		
Cash_flow					

Table 1: Classical factors

3.2 Sentiment Factors

Besides classical market or financial related factors, we also include a few sentiment factors that reflect the mood of investors and therefore may help better predict the market. We further classify the sentiment factors into two classes, the direct sentiment factors and indirect sentiment factors.

Direct Sentiment Factors

The direct sentiment factors include variables can be computed directly/extracted from Stocktwits, Sentdex(contains over 20 sources of news events) and Twitter with Quantopian. The variables include the following aspects:

Factor	description
Sentiment score	sentiment signal determined by the Sentdex algorithm, values ranging from -3 to 6.
Bullish/bearish intensity	the strength of bullishnes/bearishness on a 0-4 scale by PsychSignal's algorithm
bull_minus_bear	a net score by subtracting bearish intensity from bullish intensity
bull/bear scored messages	the total count of bullish/bearish sentiment messages
bull to bear msg ratio	ratio between bull-scored messages and bear-scored messages
No. of total messages	The number of total messages coming through

Table 2: Direct Sentiment Factors

Typical factors are shown in Table 2. Then with selected these above factors with different time frames(latest score, 3-day, 12-day, 30-day, etc.) using a simple moving average. We also weighted these variables to form combined factors(e.g the weighted average of sentiment score from Sendex and bull_minus_bear intensities from Stocktwits and Twitter). In total, we selected 19 sentiment factors that varied in time.

Indirect Sentiment Factors

The indirect sentiment factors include variables that indirectly reflect investors' mood, such as the performance of general market or economic indicators. Some of the following variables are proposed previously in literature(Dhaoui and Bensalah [2017], Chung et al. [2012], Kim and Kim [2014], Baker and Wurgler

[2006], Berger and Turtle [2012], Blanc et al. [2014]). A few financial indices may also affect the investors' opinion about the market such as price-earning ratio or market turnover rate, which we have already included them in the classical factors. Therefore, the 10 selected indirect sentiment factors are listed in Table 3.

1	number of new IPOs	6	services consumption index
2	first day IPO return	7	consumer price index
3	closed-end fund discount	8	industrial production index
4	durables consumption index	9	employment data
5	nondurables consumption index	10	dividend premium

Table 3: Indirect Sentiment Factors

4 Factor Analysis

4.1 Classical Factors

We did factor analysis for all factors through 2009 to 2017. Based on one year performance, we selected prediction factors for next year. It would be 20 classical and 10 sentiment factors with top sharp ratio every year.

We used 2017 as an example year to illustrate our factor analysis process. Firstly, we computed four measurements of each factor, including annualized return, annualized volatility, sharp ratio, max draw down.

We long the top 10% stocks, and short bottom 10% stocks in the first trading day each month according to the prediction of factors, then obtain the annualized return and annualized volatility. Instead of just buying top stocks, this strategy can lower the risk. In addition, the Sharpe ratio is the average return earned in excess of the risk-free rate per unit of volatility or total risk. We assume that free-risk rate is 2%. The max draw down is the maximum observed loss from a peak to a trough of a portfolio, before a new peak is attained. Maximum draw down is an indicator of downside risk over a specified time period.

$$sharp\ ratio = \frac{R - R_f}{\sigma}$$

$$Max\ drawdown = \frac{Trough\ value - Peak\ value}{Peak\ value}$$

As the figure *2017 factor analysis* in Appendix shows, factors have good performance within one year. Sharp ratios of 17 factors are larger than 1, and these of 7 factors are more than 2. And in 2017, we find that top3 factors in 2017 are all belongs to growth category. We sorted all factors by sharp ratio and chose top 20 as prediction factor as next year.

Factor	Return	Volatility	Sharp ratio	Max draw down
Operating income growth *Oig*	19.35%	4.71%	3.69	4.36%
Revenue growth *Rg*	32.50%	9.50%	3.21	9.33%
Sales growth *Sg*	24.69%	8.35%	2.72	8.26%
Change in shares *sharechg*	17.01%	6.45%	2.33	6.15%
PEG ratio *PEG*	16.02%	6.08%	2.31	5.75%
EPS growth *EPSG*	8.25%	2.98%	2.10	3.20%
Return over asset *ROA*	21.30%	9.58%	2.01	8.46%

Table 4: Factors in 2017 with 2+ sharp ratio

Among them, *Oig* has the top performance, and we explored it more. As the figure shows, this factor can differentiate return groups well. Long group has good performance, achieving 21.82% return, higher than

the compare basis 16.15%. By contrast, short group only has 1.73% return. We plotted the annualized return for 10 groups differentiated by this factor and obtained the same conclusion.

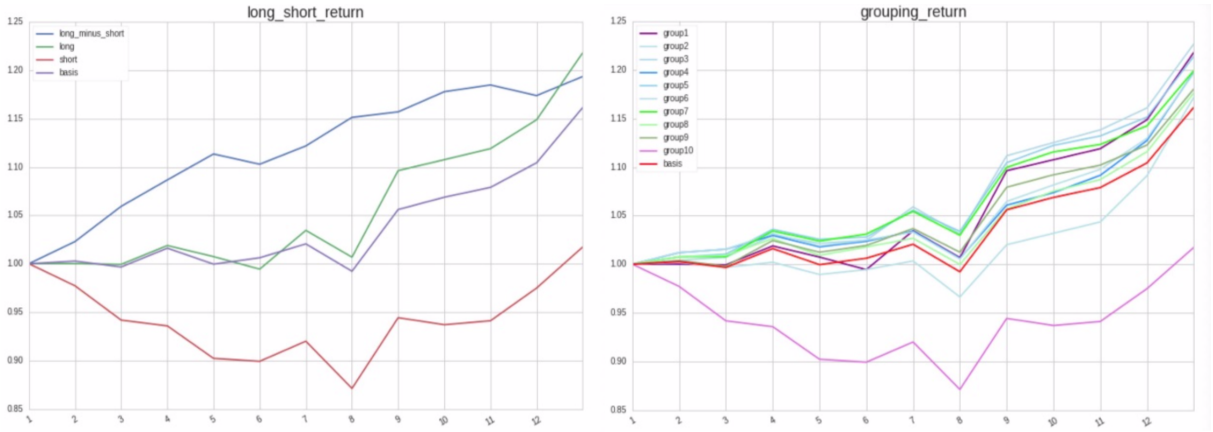


Figure 2: Results of Oig factor analysis in 2017

Then we did frequency statistics for all factors, and find that no factor can occur for every year. Net income growth and volatility factor occur 8 times in 9 years.

Index	Frequency	Index	Frequency	Index	Frequency
nig	8	mom36	5	skew	4
vol	8	sg	5	std_turn	3
bp	7	betasq	5	cf_sale	3
sharechg	7	ato	5	cr	3
mom12	7	roe	5	cf	3
size	6	roa	5	volumed	3
maxretn	6	beta	5	rg	3
epsg	6	dpsg	5	ev_ebitda	3
nig	6	bpsg	4	largretn	3
cfp	6	qr	4	peg	2
sp	6	ag	4	momchg	2
fcf	6	turn	4	ep	1
mom6	6	std_volumned	4		
Market	Momentum	Value	Growth	Profitability	Liquidity

Figure 3: Factor frequency from 2009 to 2017

After one-year factor analysis, we did it through three years, from 2015 to 2017. As the table in Appendix shows, the factor return is relatively low because factor effectiveness changes each year. Therefore, in our model we re-select our effective factors in the first month every year.

4.2 Sentiment Factors

Direct Sentiment Factors

First, we examined on Apple stock as an example to test the relationships of the stock price versus our selected direct sentiment factors. As is shown in Figure 4, it is quite obvious that a spike in the total message volume may correlate with a drop in price, since people are more reactive to a bearish market rather than a bullish market. Besides, there are correlated trends with prices and bull_minus_bear intensities though there is not very much difference between bullish and bearish intensities of the same period(the difference is normally located in the range -0.2 to 0.2). And a large increase in the bull_minus_bear intensities normally followed by an increase in the return.

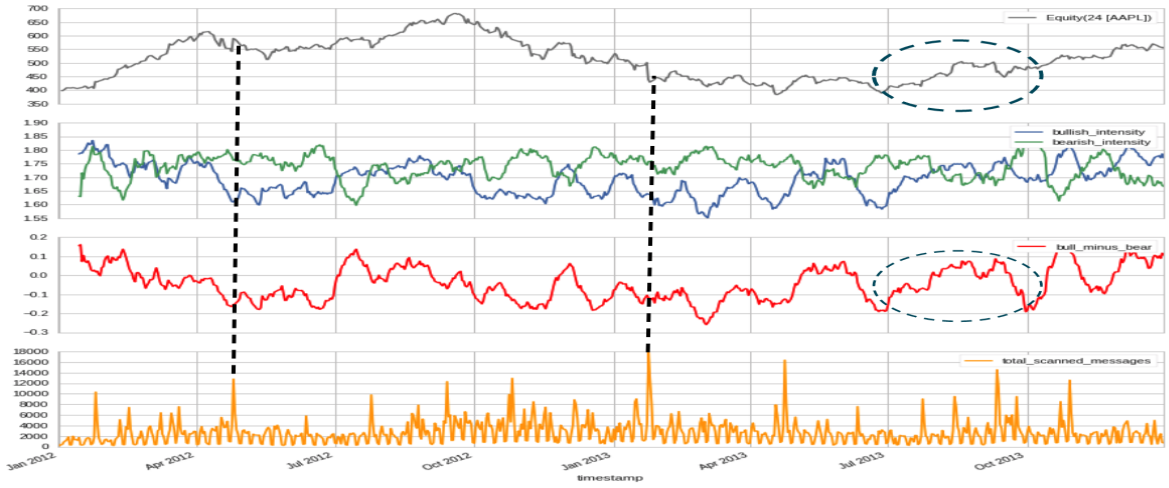


Figure 4: Direct sentiment factors example on AAPL

Secondly, we use the same method as stated in the classical factors section to form a single factor's portfolio and test its long-short Sharpe ratio from 2013 to 2017. The significant factors with Sharpe ratio greater than one are listed in Figure 5. To better analyze the results, we computed the frequencies for each factor to appear as the important factors in Figure 6. From these results we can observe that the sentiment factors are performing quite well, some of them even achieved Sharpe ratios greater than two during particular years. From the frequency table we see that the factor "Average of 3-day bullish intensity" appears to be important every year and the factor "Combined 3-day scores from Sentdex and Stocktwits" and "Number of Bull Messages" also perform quite well, appearing 4/5 years as important factors.

Therefore, we can conclude that direct sentiment factors are quite correlated with the market, the factors of the number of messages and sentiment score/intensity factors all seem significant and have a prediction power. However, similar to the classical factors, sentiment factors are also dependent on time, some factors may only be effective during one particular year. Thus we still need to re-select the factors each year. Also, a few combined factors seem to be effective as well. Time window of 3-day for computing the average scores seems to have the best performance.

Indirect Sentiment Factors

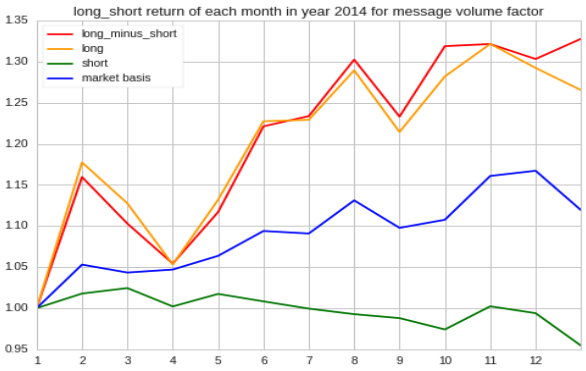
We observe the relationships of the S&P500 stock versus the selected indirect factors in Figure 7. For the 10-year period from 2009 to 2018, the following five indices seem to have a good prediction power of the returns: 1. durables consumption index, 2. nondurables consumption index, 3. services consumption index, 4. consumer price index, 5. employment data. The other factors do not possess such strong relationships.
For the five-year period from 2010 to 2015, it is shown that the industrial production index seems to have a good prediction power over this period. And for a particular time window, an increase in the number of new IPOs and their first day returns may follow by an increase in the price, while the increase in closed-end fund discount may follow by a decrease in SPY prices. These factors may not be that

2013		2014		2015		2016		2017	
Factors	Sharpe	Factors	Sharpe	Factors	Sharpe	Factors	Sharpe	Factors	Sharpe
Message volume	1.59	Bullish intensity_3	2.458	Bearish latest	3.053	Bearish score	2.130	Bearish score	1.991
Senti_return combined	1.308	Senti_sma_20	2.366	Bull_minus_bear_latest	2.418	Message volume	1.94	Bull_minus_bear_sma	1.781
Bullish intensity_3	1.279	Bearish score	2.245	Bullish_latest	1.564	Bullish score	1.846	Senti_return combined	1.654
Bull-bear-message ratio	1.124	Senti_sma_50	2.144	Bullish score	1.508	Bullish_latest	1.846	Bullish_latest	1.556
		Senti_sma_30	1.939	Bullish intensity_3	1.298	Senti_return combined	1.619	Bull messages	1.386
		Bull messages _3	1.915	Bull messages	1.288	Senti_sma_50	1.598	Bull messages_12	1.242
		Bull-bear-message ratio	1.828	Senti_latest	1.118	Bullish intensity_3	1.334	Combined sentiment	1.008
Combined sentiment	1.369	Bearish latest	1.738	Combined sentiment	1.075	Bull messages	1.303		
Bullish intensity_3	1.207	Message volume	1.696	Bull messages_12	1.04	Senti_sma_30	1.143		
Senti_latest	1.204	Bullish score	1.613			Bull_minus_bear_sma	1.112		
Bull messages	1.132	Bull messages_12	1.586			Combined sentiment	1.004		
Important factors	**4**	**Important factors**	**15**	**Important factors**	**9**	**Important factors**	**11**	**Important factors**	**7**

Figure 5: Sharpe ratios for direct sentiment factors

Index	Frequency	Index	Frequency
Average of 3-day bullish intensity	5	Average 3-day bull_minus_bear score	2
Combined 3-day scores from Sentdex and Stocktwits	4	Latest bearish score	2
Bull messages	4	Average 3-day bull-bear-message ratio	2
Total messages	3	Latest Sentdex sentiment	2
Combined 12-day bearish score from Stocktwits and twitter	3	Sentdex sentiment 30-day	2
Combined 12-day bearish score from Stocktwits and twitter	3	Sentdex sentiment 50-day	2
Bullish latest scores from Stocktwits	3	Sentdex sentiment 20-day	1
Average bull messages for 12-day	3	Bull_minus_bear_latest	1
Combined 5-day sentiment with returns	3	Average 3-day bull messages	1

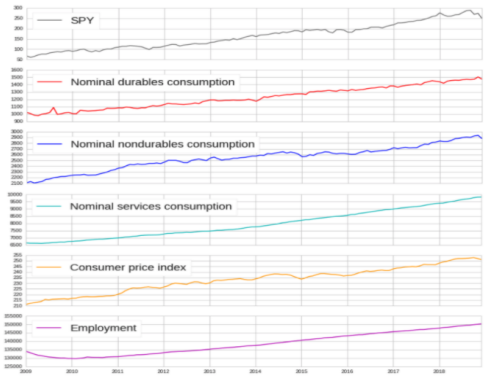
(a) Factor frequencies



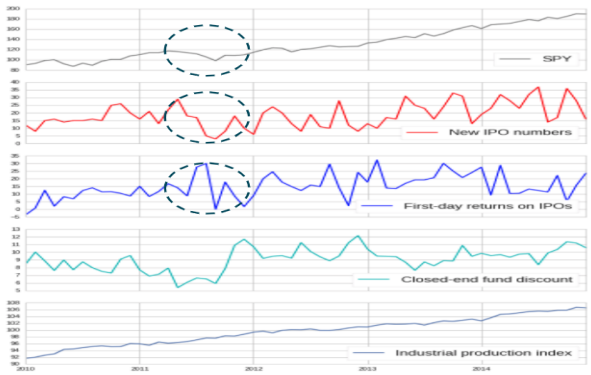
(b) Message_volume factor example

Figure 6: Direct sentiment factors portfolio test

significant compared to the first group, but we still include them in our experimented models to better test them.



(a) Indirect factors group one



(b) Indirect factors group two

Figure 7: Indirect sentiment factors

5 Experiments

5.1 Models

We use 6 different linear models as our prediction models: Ordinary least square regression, Ridge regression, Bayesian ridge regression, Lasso regression, Elastic net regression and Partial least square regression. Further, we also consider a forecasting combination model (FC) based on these 6 linear models.

5.2 Methods

Our prediction is built on a monthly rolling basis. For each year, the input factor is determined by the factor analysis of previous year, which are the factors with highest sharpe ratio of its single factor portfolio.

Then for each month of this year, the input data are the factors of each stock for the last 12 months, and the prediction target is the monthly return of current month. For example, the input data of 2013-07 is the 2012's significant factors from 2012-07 to 2013-06.

To test the performance of our prediction methods and the effective of sentiment factors, our prediction includes two parts.

5.2.1 Prediction without sentiment data

The first problem we would like to discuss is whether our monthly rolling prediction methods will work, this requires testing for a relatively longer time period. Since the sentiment data is only available after 2012, we first try prediction without sentiment data to get more data. In this part, we gather 10 years classical factors data from 2009 to 2018, then our prediction window is 9 years from 2010 to 2018. The input factors are the top 20 classical factors out of 38, updated every year.

5.2.2 Prediction with sentiment data

The second problem we want to test is the effectiveness of our sentiment factors. The sentiment data is only available after 2012 and also very limited in the first year, so in this part, we skip the 2012 and gather 6 years factors data from 2013 to 2018, then our prediction window is 5 years from 2014 to 2018. The input factors are the top 20 classical factors out of 38, top 10 direct sentiment factors out of 19, plus 10 indirect sentiment factors, updated every year.

5.3 Data Preprocessing

For each month, there are about 1000 valid stocks in our universe, so the input shape of a data point is about:$(12 * \text{average num of stocks}) * \text{num of factors}$. Then we scale the input factor data by the standard normalization method:

$$X_{\text{scale}=} \frac{X - \overline{X}}{\sigma_X}$$

Here we scale the data for the input of each month (based on the information of last 12 month) to avoid using future information. Further, to ensure the size of our dataset, we set NA values with 0 instead of deleting those data.

5.4 Portfolio Strategies

To test the performance of out prediction model in the real setting, we build our portfolio by the following strategy:

- At the start of each month, gather the factor information of last 12 months and make a prediction of each stock in our universe
- Long 5% of stocks with highest prediction return
- Short 5% of stocks with lowest prediction return
- Set a one side transaction cost of 0.25% for both long and short

6 Result

In this section, we will discuss the result of Long / Short / Long - Short / Mean portfolio with / without transaction cost.

6.1 Performance without sentiment data

For the first part of our method, we have a 9 years prediction with classical factors. Figure 4 shows the performance of 4 portfolio for OLS and Bayesian Ridge model (which has the highest return and sharpe ratio). Figure 5 shows the performance of long-short portfolio for all 7 models. Table 5 shows the statistics of long-short portfolio for all models.

We can see that the trend of all these models are very similar, this is because they are all linear models, there maybe only tiny differences between models. Bayesian ridge model has the highest annualized return of 19.95% and sharpe ratio of 1.1477, also a low max drawdown of 20.82% among these models. This result shows the power of our monthly rolling prediction method with yearly updated factors. Table 6 shows the statistics of long-short portfolio for all models with a transaction cost of 0.25% (0.5% for two sides), the highest annualized return drops to 13.05% with a sharpe ratio of 0.7062, this monthly re-balanced strategy can be largely effected by the transaction cost.

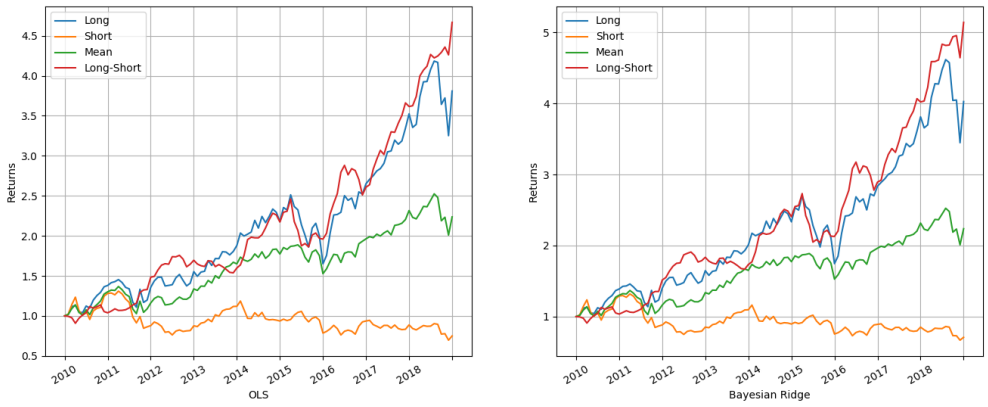


Figure 8: Portfolio performance of OLS and Bayesian Ridge

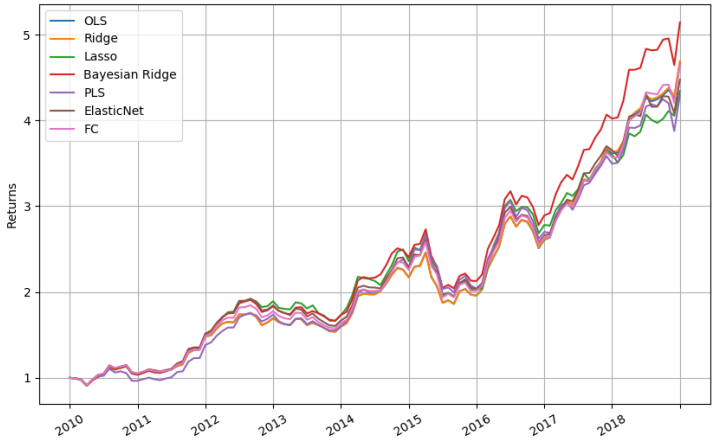


Figure 9: Portfolio performance of all 7 models

	Return	Volatility	Sharpe Ratio	Max Drawdown
Bayesian Ridge	19.95%	15.64%	1.1477	20.82%
Ridge	18.74%	14.81%	1.1305	20.75%
OLS	18.67%	14.81%	1.1252	20.75%
FC	18.59%	15.63%	1.0615	21.47%
ElasticNet	18.12%	16.15%	0.9983	21.87%
Lasso	17.73%	16.28%	0.9659	23.61%
PLS	17.58%	16.18%	0.9627	21.47%

Table 5: Statistics of long-short portfolio

	Return	Volatility	Sharpe Ratio	Max Drawdown
Bayesian Ridge	13.05%	15.64%	0.7062	20.92%
Ridge	11.90%	14.81%	0.6684	20.84%
OLS	11.83%	14.81%	0.6636	20.84%
FC	11.76%	15.63%	0.6242	21.56%
ElasticNet	11.31%	16.15%	0.5765	21.96%
Lasso	10.94%	16.28%	0.5489	23.71%
PLS	10.80%	16.18%	0.5436	21.56%

Table 6: Statistics of long-short portfolio after transaction cost

6.2 Performance with sentiment data

In the second part of our method, we have a 5 years prediction with sentiment factors. To test the effectiveness of sentiment data, we compare the performance of our models with / without sentiment factors. Figure 6 shows the performance of 4 portfolio for OLS model with and without sentiment factors (which is one of the highest return and sharpe ratio). Figure 7 shows the performance of long-short portfolio for all 7 models. Table 7 and 8 shows the statistics of long-short portfolio for all models without / with sentiment data.

We can see that the trend of all these models are still very similar, but there are more differences between models than the previous result. For all models, we get improvements in every aspect by adding sentiment data. For example, OLS model has the highest sharpe ratio of 1.2832 without sentiment factors, also a high return of 25.57% and the lowest max drawdown of 22.54% among these models. After adding sentiment factors, the sharpe ratio increases to 1.3557 with a higher return of 26.11%. This result shows the effectiveness of sentiment data.

Table 9 and 10 shows the statistics of long-short portfolio for all models without / with sentiment data after considering a transaction cost of 0.25% (0.5% for two sides). The annualized return of OLS without sentiment factors drops to 18.36% with a sharpe ratio of 0.8909. The annualized return of Elastic Net with sentiment factors drops to 20.43% with a sharpe ratio of 0.9903. Even though there is a significant drop of performance for all models, the portfolio with sharpe ratio about 1 is still very satisfactory. This shows the potential of applying sentiment data into factor model.

Figure 10: Portfolio performance of OLS

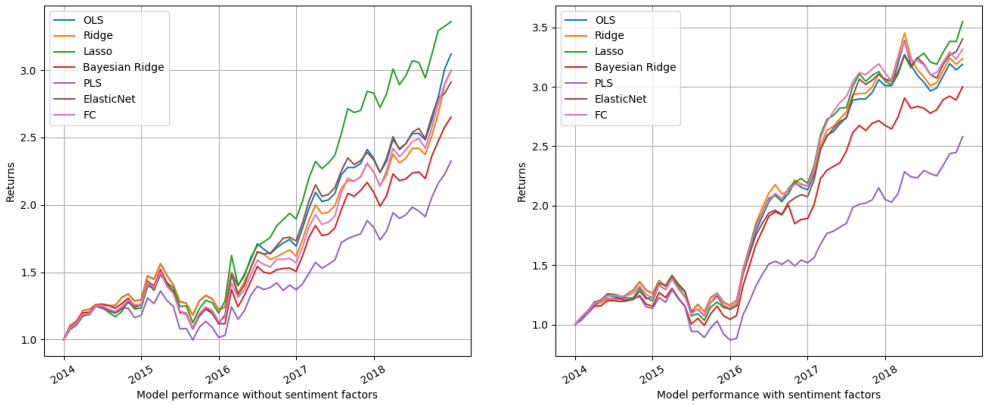


Figure 11: Portfolio performance of all 7 models

	Return	Volatility	Sharpe Ratio	Max Drawdown
OLS	25.57%	**18.36%**	**1.2832**	**22.54%**
Lasso	**27.44%**	20.86%	1.2194	31.44%
Ridge	24.56%	18.91%	1.1930	25.97%
FC	24.57%	19.46%	1.1602	24.01%
ElasticNet	23.83%	20.16%	1.0827	28.26%
Bayesian Ridge	21.54%	20.05%	0.9749	26.11%
PLS	18.39%	18.72%	0.8756	23.13%

Table 7: Statistics of long-short portfolio without sentiment factors

	Return	Volatility	Sharpe Ratio	Max Drawdown
ElasticNet	27.75%	18.61%	**1.3836**	21.19%
OLS	26.11%	**17.78%**	1.3557	**21.11%**
Lasso	**28.83%**	19.83%	1.3534	24.09%
FC	27.11%	18.93%	1.3263	22.88%
Ridge	26.48%	18.64%	1.3137	23.38%
Bayesian Ridge	24.58%	19.52%	1.1565	25.33%
PLS	20.87%	20.24%	0.9323	24.68%

Table 8: Statistics of long-short portfolio with sentiment factors

	Return	Volatility	Sharpe Ratio	Max Drawdown
OLS	18.36%	**18.36%**	**0.8909**	**22.63%**
Lasso	**20.13%**	20.86%	0.8692	31.57%
Ridge	17.40%	18.91%	0.8147	26.08%
FC	17.42%	19.46%	0.7925	24.11%
ElasticNet	16.72%	20.16%	0.7299	28.37%
Bayesian Ridge	14.55%	20.05%	0.6259	26.22%
PLS	11.57%	18.72%	0.5110	23.22%

Table 9: Statistics of long-short portfolio without sentiment factors after transaction cost

	Return	Volatility	Sharpe Ratio	Max Drawdown
ElasticNet	20.43%	18.61%	**0.9903**	21.28%
Lasso	**21.46%**	19.83%	0.9812	24.19%
OLS	18.87%	**17.78%**	0.9489	**21.20%**
FC	19.82%	18.93%	0.9414	22.97%
Ridge	19.23%	18.64%	0.9245	23.47%
Bayesian Ridge	17.42%	19.52%	0.7900	25.43%
PLS	13.91%	20.24%	0.5885	24.79%

Table 10: Statistics of long-short portfolio with sentiment factors after transaction cost

7 Conclusion and Future Work

In summary, we have shown that the factors are time dependent and may only be effective during specific time period and they may have a better performance in short period than a longer period, so it is a good idea to consider a prediction method of re-selecting significant factors every year in factor models.

Besides, we show the potential of of applying sentiment data into factor model. By adding those sentiment data to the factor pool, we can get a significant improvement on the model performance.

In the future, several ways we may try as the next step:

- Due the limitation of Quantopian platform, we cannot apply deep learning networks such as RNN or TCN(temporal convolutional networks). Those non-linear advanced machine learning models may have a better prediction power for this problem.
- In this project, we just use some sentiment data directly from Quantopian. We can apply NLP techniques to compute more well designed sentiment signals based on lexicons/bi-grams/n-grams.
- Here we basically use sharpe ratio to measure the performance of single factor portfolio and select effective factors. We can use more measurements to evaluate factors and try to create some compound factors with better performance.
- The correlation between factors can also be explored in detail.

8 Appendix

Factor	Description	Factor	Description
Size	Market capitalization	Beta, Betasq	Market beta and its square
Vol	The volatility of 1y stock price	Skew	The skewness of 1y stock price
Turn, std_turn	Stock 1y turnover and its volatility	Volume, Volume_std	trading volume and its volatility
Maxetn	Maximum daily return in 1y	Sharechg	Changes in shares outstanding
Cash_flow	Cash flow in quarterly report		

Table 11: Market factors

Factor	Description		
Mom6/12/36	Cumulative return in past 6/12/36 months	Lagretn	Short term reversal, the return last month
Momchg	Cumulative return in past 6 months minus that between past 6 to 12 months		

Table 12: Momentum factors

Factor	Description	Factor	Description
BP	Book-to-price ratio	EP	Earnings-to-price ratio
CFP	Cash flow to price ratio	SP	Sales-to-price ratio
PEG	Price/Earnings-to-Growth	EV_Ebitsa	Enterprise value-to-EBITDA

Table 13: Value factors

Factor	Description	Factor	Description
BP	Book-to-price ratio	EP	Earnings-to-price ratio
CFP	Cash flow to price ratio	SP	Sales-to-price ratio
PEG	Price/Earnings-to-Growth	EV_Ebitsa	Enterprise value-to-EBITDA

Table 14: Growth factors

Factor	Description	Factor	Description
ROA	Return over asset in quarterly report	ROE	Return over equity in quarterly report
ATO	Asset turnover in quarterly report	FCF_Yield	Free cash flow yield

Table 15: Profitability factors

Factor	Description	Factor	Description
CR	Current ratio in quarterly report	QR	Quick ratio in quarterly report
CF_sale	Cash flow to sales		

Table 16: Liquidity factors

	annualized return	annualized volatility	sharpe ratio	max_drawdown
oig	0.193533	0.047072	3.686569	0.043564
rg	0.324957	0.094977	3.210854	0.093339
sg	0.246889	0.083456	2.718671	0.082583
sharechg	0.170140	0.064512	2.327301	0.061537
peg	0.160150	0.060752	2.306928	0.057520
epsg	0.082459	0.029796	2.096240	0.032018
roa	0.213041	0.095836	2.014285	0.084363
roe	0.189501	0.085986	1.971258	0.081629
maxretn	0.268326	0.134365	1.848144	0.108884
nig	0.100179	0.044602	1.797649	0.045268
qr	0.102355	0.049452	1.665346	0.034267
ev_ebitda	0.126239	0.072197	1.471514	0.066291
ag	0.136120	0.087147	1.332468	0.080595
turn	0.124588	0.084796	1.233398	0.089538
cr	0.091463	0.062172	1.149437	0.042726
cfp	0.144183	0.118451	1.048397	0.074747
ato	0.114848	0.091440	1.037265	0.085767

	annualized return	annualized volatility	sharpe ratio	max_drawdown
qr	0.059021	0.072818	0.535866	0.095738
fcf	0.068379	0.090907	0.532184	0.107813
cf_sale	0.079116	0.117348	0.503766	0.168466
sharechg	0.054280	0.082737	0.414329	0.113644
ato	0.065946	0.138383	0.332019	0.159320
bpsg	0.048453	0.088065	0.323094	0.095738
ep	0.049834	0.093012	0.320753	0.101733
cfp	0.069245	0.176185	0.279511	0.197651
rg	0.060415	0.150469	0.268596	0.180635
oig	0.052445	0.131491	0.246746	0.183852
sg	0.054703	0.143238	0.242276	0.170606
peg	0.041391	0.107121	0.199694	0.120892
ag	0.045382	0.129910	0.195380	0.153302
volumned	0.040540	0.118327	0.173583	0.155849
roa	0.046543	0.153461	0.172966	0.225822
std_volumned	0.039040	0.114573	0.166178	0.153490
size	0.047321	0.166024	0.164563	0.232344

Figure 12: Results of factor analysis in 2017, and through 2015 to 2017

Acknowledgments

The authors would like to thank Enguerrand Horel, and Dr. Lisa Borland for their helpful suggestions and feedback for this project in the whole quarter.

References

Malcolm Baker and Jeffrey Wurgler. Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4):1645–1680, 2006.

Dave Berger and HJ Turtle. Cross-sectional performance and investor sentiment in a multiple risk factor model. *Journal of Banking & Finance*, 36(4):1107–1121, 2012.

Alexios Beveratos, Jean-Philippe Bouchaud, Stefano Ciliberti, Laurent Laloux, Yves Lempérière, Marc Potters, and Guillaume Simon. Deconstructing the low-vol anomaly. *The Journal of Portfolio Management*, 44(1):91–103, 2017.

Pierre Blanc, Rémy Chicheportiche, and Jean-Philippe Bouchaud. The fine structure of volatility feedback ii: Overnight and intra-day effects. *Physica A: Statistical Mechanics and its Applications*, 402:58–75, 2014.

Jean-Philippe Bouchaud, Stefano Ciliberti, Augustin Landier, Guillaume Simon, and David Thesmar. The excess returns of" quality" stocks: a behavioral anomaly. *arXiv preprint arXiv:1601.04478*, 2016.

Mark M Carhart. On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82, 1997.

San-Lin Chung, Chi-Hsiou Hung, and Chung-Ying Yeh. When does investor sentiment predict stock returns? *Journal of Empirical Finance*, 19(2):217–240, 2012.

Stefano Ciliberti, Emmanuel Sérié, Guillaume Simon, Yves Lempérière, and Jean-Philippe Bouchaud. The size premium in equity markets: Where is the risk? *The Journal of Portfolio Management*, 45(5): 58–68, 2019.

Abderrazak Dhaoui and Nesrine Bensalah. Asset valuation impact of investor sentiment: A revised fama–french five-factor model. *Journal of Asset Management*, 18(1):16–28, 2017.

Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of*, 1993.

Soon-Ho Kim and Dongcheol Kim. Investor sentiment from internet message postings and the predictability of stock returns. *Journal of Economic Behavior & Organization*, 107:708–729, 2014.