CUFE·CAFD

TITLE： BIG DATA FINAL PROJECT

NAME： 孟舒晨 2019212176

DATE： 01/09/2020

1. Introduction

Does market sentiment affect stock picking? A large number of recent researches have focus on this question. Market sentiment always be concerned as the related factors to business cycles and financial crises. The market sentiment index is formed by the news in this paper and assign financial market sentiment on a stock playing a central role in predicting the return of this stock.

In this article, machine learning models are used. In these models, we take a different approach to predict the stocks' performance. We can divide this process into discrete process and continuous process. As for discrete process, the training data is seven factors value at the day of 1 months before, and the training target is binary label (better than bench mark or not). Test data includes seven factors value today, the predicted probability of each stock beating bench mark in the testing set and rank the probability and select top10/20. As for continuous process, the training data is seven factors value at the day of 1 months before, and the training target is continuous label (predicted return of each stock). Test data includes seven factors value today, the predicted return of each stock in the testing set and rank the return and select top10/20.

Consistent with this hypothesis, we document that factors, especially quantity factors that have previously been shown to forecast returns in the stock market also have significant predictive power. The factor selection follows "*Quality Minus Junk*" (Asness, Frazzini, and Pedersen (2019))

This paper using data from UQER, Tushare, over the period from 2016.1.1 to

2019.12.1, indicates that emotion factors really make sense. I have tried the basic machine learning methods without emotion factor, and the results shows that the average annualized return is about 5%, which indicates emotion factor did increase the annuity return. At first, I want to compared these three methods with each other, and find the best one, and analysis the reasons, however, randomness was found in the back test results. In different back-testing period, we will have different back-testing results.

In section 3, my empirical analysis is organized as follows. First, I begin by using broadcast texts from CCTV news, and forming keywords for emotional analysis as the indicator of overall position control. Then, combined with the percentage of related news in the total news of the day and news emotion, a news popularity index is formed and updates daily. In the following parts, I dig deeper into the nature of the stocks prices and try to select stocks by using machine learning methods. Factors, including emotion factor, are used as characteristics, and three different machine learning methods are used to select stocks, and their performance was compared. In the end, The back-testing results of several strategies are shown in the section 4. My technical roadmap is shown below:

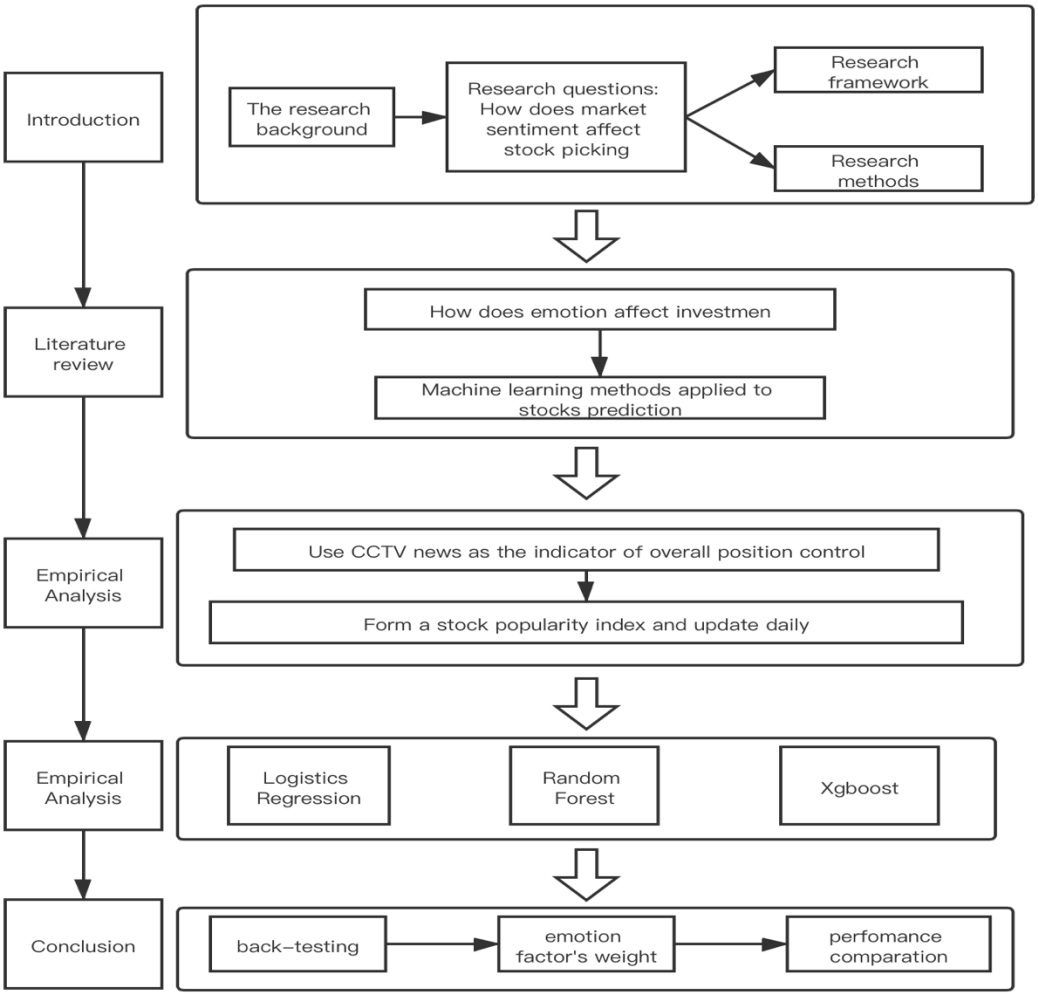


Figure 1: Method map

In addition, if fluctuations in market sentiment are causing movements in the price of stock, my methodology should cover the fluctuation in stock market sentiment.

2. Literature review

The study focuses on the influence of emotion starting from Campbell and Shiller (1988), who says that the emotional state of investors when they decide on their investment is no doubt one of the most important factors causing the bull market. Tetlock (2007) quantifies the interaction between media and the stock market with the daily contents of the Wall Street journal column. They find that the high degree of pessimism in the media predicted downward pressure on market prices and then returned to fundamentals, with abnormally high or low pessimism predicting high market volume. Odean (1998) gives an empirical evidence by investigating almost 100000 transaction by retailer investor during 1987-1993, and finds out that the frequency of winner/loser sales relative to the opportunities for winner/loser sales, compare the actual sale of winner/loser with sales that could have been made at a gain/losses(avoid). López-Salido, Stein, and Zakrajšek (2017) use U.S. data from 1929 to 2015, show that elevated credit-market sentiment in year t − 2 is associated with a decline in economic activity in years t and t + 1. In particular, Greenwood and Hanson (2013) have shown that when the credit spread of corporate bonds was narrow relative to the historical normal level, and when the proportion of high-yield (or "junk") bond issuance in the total amount of corporate bond issuance rose, this often indicated that the return of credit investors would decrease in the future. These papers show that market sentiment does affect stock returns.

Brown and Cliff (2004) investigate investor sentiment and its relation to near-term stock market returns. They also find that many commonly cited indirect

measures of sentiment are related to direct measures (surveys) of investor sentiment. Baker and Wurgler (2007) It shows that it is possible to measure investor sentiment and that mood swings have significant, important and regular effects on individual companies and the stock market as a whole. Mittal and Goel n.d.(2017) using the principles of emotion analysis and machine learning, find the correlation between "public sentiment" and "market sentiment". They use twitter data to predict public sentiment, and use the predicted sentiment and the value of the Dow-Jones industrial average (DJIA) in previous days to predict stock market movements. These papers show the proxy variables of emotions and how emotions affect the stock market.

As we can see, traditional methods are not suitable for this question, machine learning methods have been widely applied in this problem, and the sources of data have been diversity, like media and Twitter. As for machine learning, Zhou et al. (2019)attempts to establish a learning structure LR2GBDT for stock index prediction and trading mainly through logistic regression. It not only performs better than other models, but also has significant improvement in statistics and economy. It can make use of simple trading strategies, even taking into account transaction costs. A new method is proposed by Khaidem, Saha, and Dey (2016) to minimize the risk of investing in the stock market by predicting stock returns using a powerful machine learning algorithm called integrated learning. The learning model used is a set of multiple decision trees. The results show that the performance of this algorithm is better than the existing algorithms. A new sparse sensing algorithm and weighted quantile sketch approximate tree learning algorithm are proposed by Chen and

Guestrin (2016). More importantly, they provide insights into cache access patterns, data compression, and shadows to build scalable tree enhancement systems. Combined with these insights, billions of examples can be done with far fewer resources than existing systems.

3. Data and Empirical Analysis

CCTC News

Data loading: There are two ways to obtain the CCTV news broadcast text of the past 10 years. One is to write the crawler by myself and crawl the news broadcast from CCTV website. The other is to get the data for free through the API of Tushare SDK. In this paper, I loaded these data though the API of Tushare SDK.

date	title	content
20181222	在新时代创造中华民族新的更大奇迹——习近平总书记在庆祝改革开	习近平总书记在庆祝改革开放40周年大会上的重要讲话，发出了新
20181222	坚定发展信心 坚持稳中求进——中央经济工作会议精神在全国干部群	昨天闭幕的中央经济工作会议，总结2018年经济工作，分析当前经
20181222	人民日报评论员文章：为全面建成小康社会收官打下决定性基础——	明天出版的人民日报将发表评论员文章，题目是《为全面建成小康社
20181222	李克强签署国务院令 公布修订后的《中华人民共和国个人所得税法实	国务院总理李克强日前签署国务院令，公布修订后的《中华人民共和
20181222	全国政协召开双周协商座谈会 围绕"推进境外经贸合作区建设"建言资	十三届全国政协第十七次双周协商座谈会近日在京召开。中共中央政
20181222	【坚持高质量发展笃定前行】生态文明建设推动共建美丽中国	生态优先、绿色发展的理念正在助力美丽中国建设，实现经济高质量
20181222	人民日报评论员文章：我们为创造奇迹的中国人民感到无比自豪——	明天出版的人民日报将发表评论员文章，题目是《我们为创造奇迹的
20181222	改革先锋风采	改革开放40年先锋人物，今天关注中国天眼的主要发起者及奠基人
20181222	尤权受中共中央委托向党外人士通报中央经济工作会议精神	12月22日，受中共中央委托，中共中央书记处书记、中央统战部部
20181222	我国首颗低轨宽带通信技术验证卫星发射成功	今天上午7时51分，我国首颗低轨宽带通信技术验证卫星在酒泉卫星
20181222	广西：从"路网末梢"跻身"区域枢纽"	广西作为面向东盟开放的"桥头堡"，五年来高铁从无到有，营运里程
20181222	【为了民族复兴·英雄烈士谱】杨石魂：坚贞不屈中华魂	《为了民族复兴·英雄烈士谱》系列报道，今天为您讲述杨石魂的革
20181222	国内联播快讯	田湾核电二期工程全面建成投产今天（22日），田湾核电站4号机组
20181222	亚投行和新开发银行成为联大观察员	第73届联合国大会20号协商一致通过决议，邀请亚洲基础设施投资
20181222	伊朗开始军演 美航母进入波斯湾水域	伊朗伊斯兰革命卫队今天（22号）在波斯湾举行大规模海岸炮演
20181222	国际联播快讯	土推迟对叙库尔德武装军事行动土耳其总统埃尔多安21日表示，经

Then with the Python package, jieba, as a tool for word segmentation, we can quickly implement text words segmentation. At the same time, we can set the keyword black list and white list, filter the unnecessary, extract the desired keywords. After processing word segmentation by date, we can make statistics on word frequency and generate a complete words frequency CSV file. And to make it more observable, I use wordcloud api to plot wordcloud with China map as background and use matplotlib animation function to generate animating wordcloud visualization of CCTV news. Use this word frequency CSV file, and some defined rules, we can analyze the emotion of news broadcast after segmentation, judge whether the news is positive or negative. And finally, we can use this information to calculate the

percentage of positive news that day. The two animation examples of CCTV news word-cloud visualization could be found here:

https://mengmeng12.github.io/blog/wc_animation.mp4

https://mengmeng12.github.io/blog/wc_animation1.mp4

News emotion factor

The UQER news heat index is used to obtain the news heat index of the security for a period of time (that is, the percentage of the number of related news today of this security in the total news of the day), and the data is updated daily. The critical parameters to obtain the data are secID, beginDate, and endDate.

Then we do some emotion analysis, in the data list obtained, each row is the corresponding heat Index of securities on a certain day, with negative numbers

representing negative emotions and positive numbers representing positive emotions. Here is an example:

Table 1: Sentiment index

secID	exchangeCD	exchangeName	ticker	secShortName	sentimentIndex
000831.XSHE	XSHE	深圳证券交易所	831	五矿稀土	-0.294702
600030.XSHG	XSHG	上海证券交易所	600030	中信证券	-0.171486
600489.XSHG	XSHG	上海证券交易所	600489	中金黄金	-0.162101
601225.XSHG	XSHG	上海证券交易所	601225	陕西煤业	-0.162101
002653.XSHE	XSHE	深圳证券交易所	2653	海思科	-0.15472

Stock prediction process

Logistic regression

According to the idea of multiple linear regression, we could first start to linear regression, what many students do in the mid-term project, many students applied the multiple factor regression strategy, according to the factor loading to choose stocks, but this method requires all the factors of value and benefits are not particularly unusual, if we are in the case of the extreme situation, then, the influence of the outliers on the regression results will be bigger, to avoid this situation, we chose the sigmod function to help us solve this problem.

$$P = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}$$

$$Q = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

$$\text{logit}P = \ln\frac{P}{Q} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Because of the characteristics of the function, the output result is no longer the prediction result, but the probability that a value is predicted as a positive example.

$$P_i = \frac{e^{\beta_0+\beta_1 x_i+\beta_2 x_{i_2}+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_{i_1}+\beta_2 x_{i_1}+\cdots+\beta_p x_p}} \qquad p_i(y_i=1)$$

$$Q_i = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_1+\cdots+\beta_p x_p}} \qquad p_i(y_i=0)$$

The selection of threshold value is very important. We use the maximum likelihood estimation method to maximize the probability that all the predicted results are correct.

$$L = \prod_{i=1}^{n} l_i = \prod_{i=1}^{n} P_i^{y_i} Q_i^{1-y_i}$$

$$\ln(L) = \sum_{i=1}^{n} [y_i \ln P_i + (1-y_i)\ln Q_i]$$

$$\ln(L) \to max \qquad B^* = (b_0, b_1, \ldots\ldots, b_k)$$

Random forest

Another way to predict is random forest. There are many classification trees in a random forest. We're going to classify a test sample, classification depends on the votes cast. Each tree in the forest is independent, and the 99.9 percent of unrelated trees make predictions that will cancel each other out. A few good trees will make a good prediction beyond the noise of the crowd.

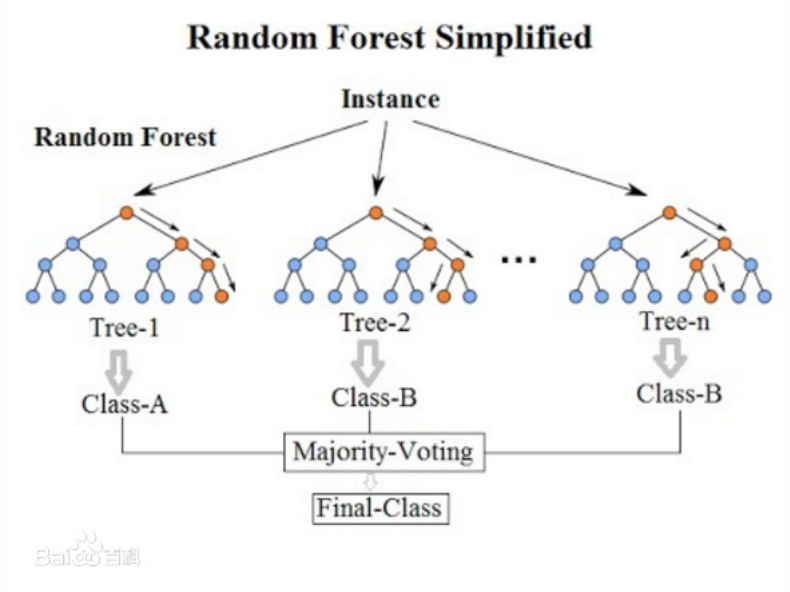


Figure 2: Random Forest

The idea of bagging in random forest is to vote the classification results of several weak classifiers to form a strong classifier.

$$\mathrm{I}(s_1, s_2, \cdots\cdots, s_m) = \sum p_i \log 2\,(p_i)(\mathrm{i} = 1..\mathrm{m})$$

I(x) is the information for the random variable, and p(xi) is the probability when xi occurs. This is the index used to select the feature, and the smaller the I, the better the selectivity of the feature. There are two kinds of combined strategies, one is the average method and the other is the voting method. Here, the simple average and the majority bidding method are shown:

$$H(x) = \frac{1}{T}\sum_{i=1}^{T} h_i(x)$$

$$H(x) = \begin{cases} c_j, & \text{if } \sum_{i=1}^{T} h_i^j(x) > 0.5 \sum_{k=1}^{T}\sum_{i=1}^{T} h_i^k(x) \\ \text{reject}, & \text{otherwise.} \end{cases}$$

XG-Boost

XG-Boost is one of the boosting method. The idea of Boosting is to integrate many weak classifiers together to form a strong classifier. The tree model used is the CART regression tree model. CART regression tree is assumed to be a binary tree by splitting the features. For example, the current tree node is divided based on the eigenvalue.

$$R_1(j,s) = \{x|x^{(j)} \leq s\} \; and \; R_2(j,s) = \{x|x^{(j)} > s\}$$

Suppose that samples with the eigenvalue less than s are divided into left subtrees, and samples with the eigenvalue greater than s are divided into right subtrees.

$$\sum_{i \in R_m} (y_i - f(x_i))^2$$

The optimal function for this partition：

$$\min_{j,s}\left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2\right]$$

$$\hat{y} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i)$$

where $F = \{f(x) = w_{q(x)}\}(q: R^m \to T, w \in R^T)$

The idea is to keep adding trees, keep doing feature splitting to grow a tree, adding one tree at a time is actually learning a new function to fit the residual predicted last time. When we finish the training and get k trees, we need to predict the score of a sample.

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

The newly generated tree is to fit the residual of the previous prediction, that is, when t trees are generated, the prediction fraction can be written as follows:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t)$$

$$w_j^* = -\frac{G_j}{H_j + \lambda} Obj = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$

In fact, according to the characteristics of the sample, a corresponding leaf node will fall in each tree, and each leaf node will correspond to a score. Finally, we just need to add up the corresponding score of each tree to be the predicted value of the sample.

Stock prediction process

We can divide this process into discrete process and continuous process. As for discrete process, the training data is seven factors value at the day of 1 months before, and the training target is binary label (better than bench mark or not). Test data includes seven factors value today, the predicted probability of each stock beating bench mark in the testing set and rank the probability and select top10/20. The factors are shown below:

Table 2: Factor discrimination

Classification	Factor
Profitability	Return on equity
	Price Earnings
	Return On Assets
Volitilaty	Volitilaty
Market	beta
Sentiment	sentiment index
	heat index

* The factor selection follows "*Quality Minus Junk*" (Asness, C. S., Frazzini, A., and Pedersen,

L.H., (2017)).

As for continuous process, the training data is seven factors value at the day of 1 months before, and the training target is continuous label (predicted return of each stock). Test data includes seven factors value today, the predicted return of each stock in the testing set and rank the return and select top10/20.

4. Conclusion

Characteristics of CCTV-News

The picture below shows the percentage of positive deviation from the mean.

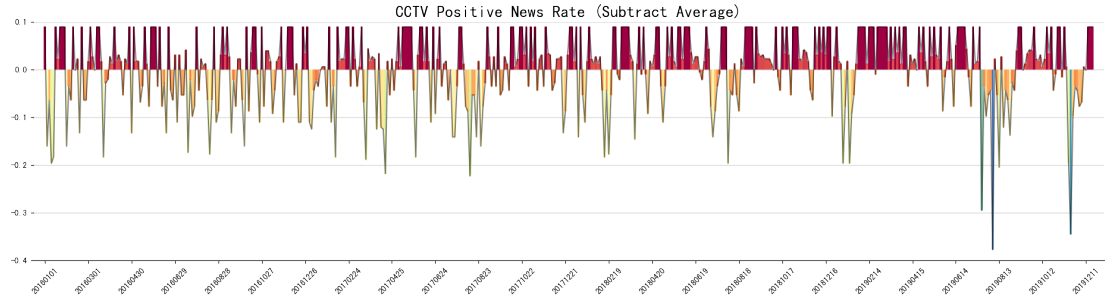


Figure 3: CCTV Positive News Rate (Subtract Average)

I adopt the deviations from the cumulative changes (if it is bigger than 1, then it equals to 1) to determine the change of position, to avoid over reaction.

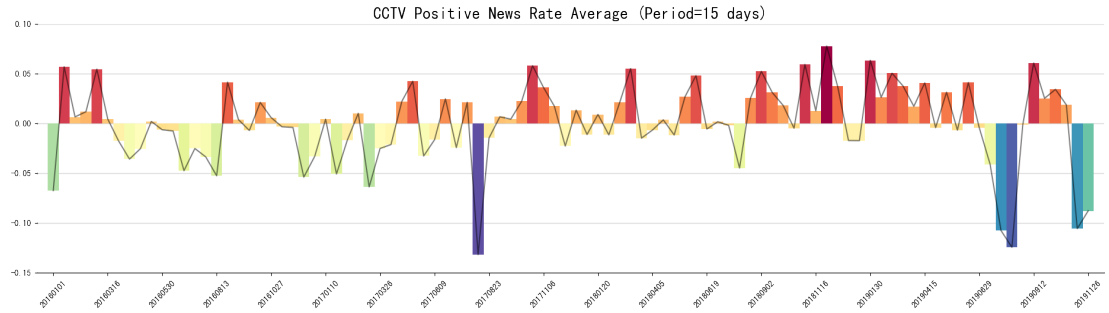


Figure 4: CCTV Cumulative Positive News Rate (Subtract Average)

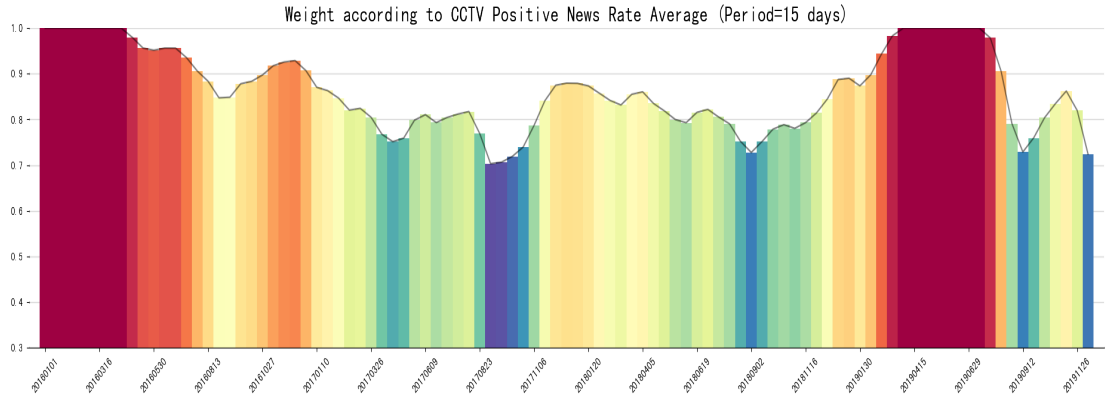


Figure 5: Weight according to CCTV News Rate Average (Period=15 days)

The figure above shows the adjusted total positions according to CCTV news, as we can see, the percentage of positive news of CCTV is around 0.9 most of the time,

but the percentage on different dates is slightly different.

Simple strategy using emotion factor

Then I tried the simple strategy using emotion factor, that is, buy stocks whose market news is hot and news sentiment is positive. The strategy's parameters is shown below:

Start date: January 1, 2016

End date: December 1, 2019

Adjustment cycle: 30 trading days

Buy method: equal weight buy

Rule: pick the 100 stocks with the highest heat, and then pick the 20 stocks with the highest emotion and the highest positivity.

The performance of this strategy is shown below:

Table 3:Back-testing indicators

Indicator			
Annualized Return	-9.90%	**sharpe ratio**	-0.56
CSI300 Annualized Return	1.90%	**volatility**	22.70%
α	-11.00%	**information ratio**	-1.11
β	1.08	**Maximum Drawdown**	37.00%

Cumulative return

Since this strategy of picking stocks on a hot spot is usually applied by the retail investors, it turns out that retail investors always lose money.

Machine learning results

The machine learning strategies can be summarized as below, so I won't repeat about this from one method to another.

Strategy: select best 10/20 stocks (by quantifying probability to beat benchmark). sell worst 10/20.

Buy method: equal weight buy(but the total positions are adjusted (by CCTV news emotional index).

This method enhances the effectiveness and efficiency of multi factor stock selection model by its accuracy in stock classification forecast.

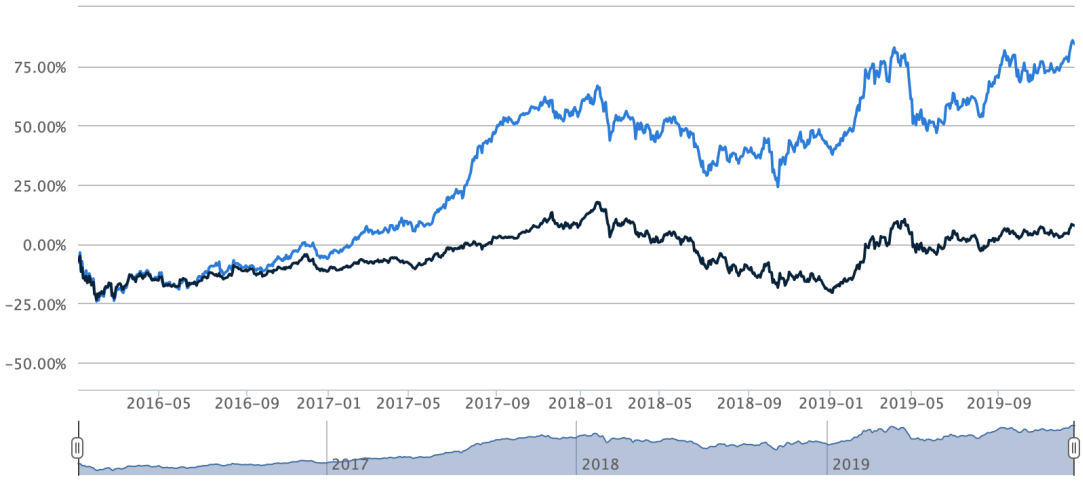
Method 1:

Table 4: Back-testing indicators

Indicator			
Annualized Return	17.10%	**sharpe ratio**	0.62
CSI300 Annualized Return	1.90%	**volatility**	21.90%
α	15.20%	**information ratio**	1.44

β	1.02	**Maximum Drawdown**	25.60%

Cumulative return



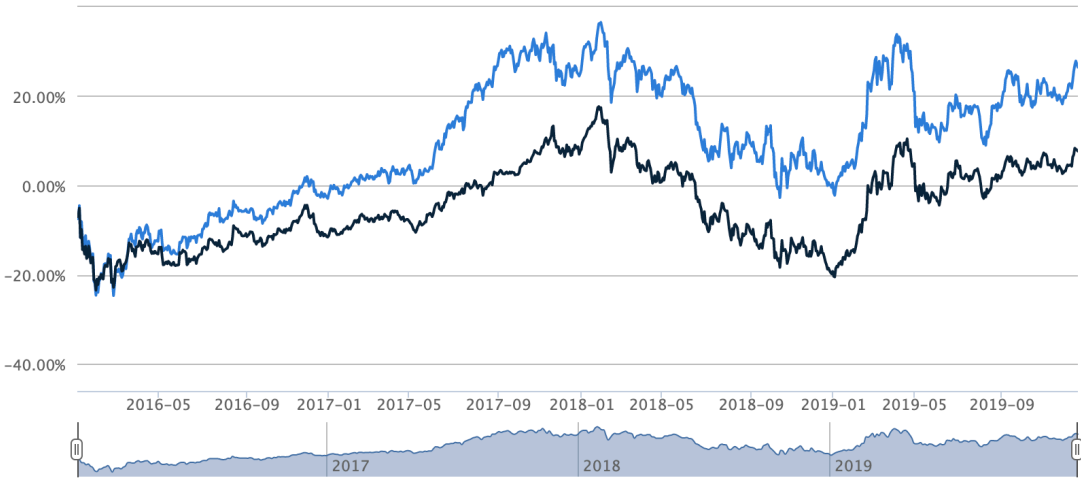
As we can see, when we add quality factors into our strategy and using logistics to pick up tocks, the annualized return is increased from -9.9% to 17.1%, which is a huge progress. But as for risk indicators, sharpe ration and maximum drawdown does not perform good, they are 0.62 and 25.6%, but the information ratio is bigger than 1, indicates that when we gain the same return, we suffer smaller risk compared with benchmark.

Method 2:

Table 5: Back-testing indicators

Indicator			
Annualized Return	6.30%	**sharpe ratio**	0.12
CSI300 Annualized Return	1.90%	**volatility**	22.30%
α	4.40%	**information ratio**	0.51
β	1.06	**Maximum Drawdown**	28.80%

Cumulative return



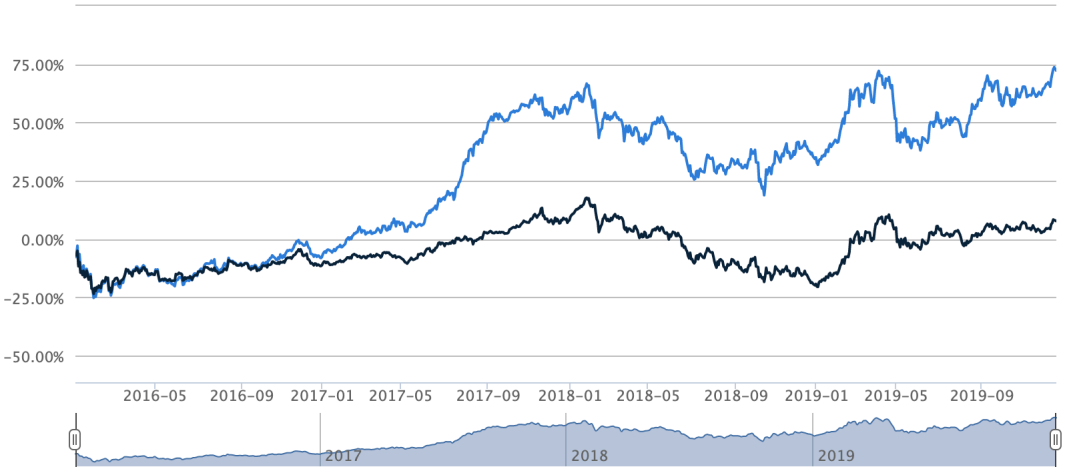
As we can see, when we add quality factors into our strategy and using random forest to pick up tocks, the annualized return is increased from -9.9% to 6.3%, which does not perform as good as logistics. And as for risk indicators, sharpe ration and maximum drawdown does not perform good, they are 0.12 and 28.8%, and the information ratio is smaller than 1, indicates that when we gain the same return, we suffer bigger risk compared with benchmark.

Method 3（discrete）：

Table 6: Back-testing indicators

Indicator			
Annualized Return	15.10%	**sharpe ratio**	0.52
CSI300 Annualized Return	1.90%	**volatility**	22.10%
α	13.20%	**information ratio**	1.24
β	1.03	**Maximum Drawdown**	28.70%

Cumulative return



As we can see, when we add quality factors into our strategy and using XG-Boost to pick up tocks, the annualized return is increased from -9.9% to 15.1%, which is a huge progress. And this method has the smallest volatility. Sharpe ration and maximum drawdown does not perform good, they are 0.52 and 28.7%, but the information ratio is bigger than 1, indicates that when we gain the same return, we suffer smaller risk compared with benchmark.

Method 3（continuous）：

Table 7: Back-testing indicators

Indicator			
Annualized Return	19.00%	**sharp ratio**	0.65
CSI300 Annualized Return	1.90%	**volatility**	23.70%
α	17.10%	**information ratio**	1.21
β	1.02	**Maximum Drawdown**	33.10%

This is the best result, the annualized return is increased from -9.9% to 19.0%, which is a huge progress. And the information ratio is bigger than 1, indicates that when we gain the same return, we suffer smaller risk compared with benchmark.

These results show that machine learning methods enhance the effectiveness and efficiency of multi factor stock selection model by its accuracy in stock classification forecast. And in the back-testing period (2016.1.1-2019.12.1), the XG-Boost method's performance is the best.

Machine learning results——factor weights

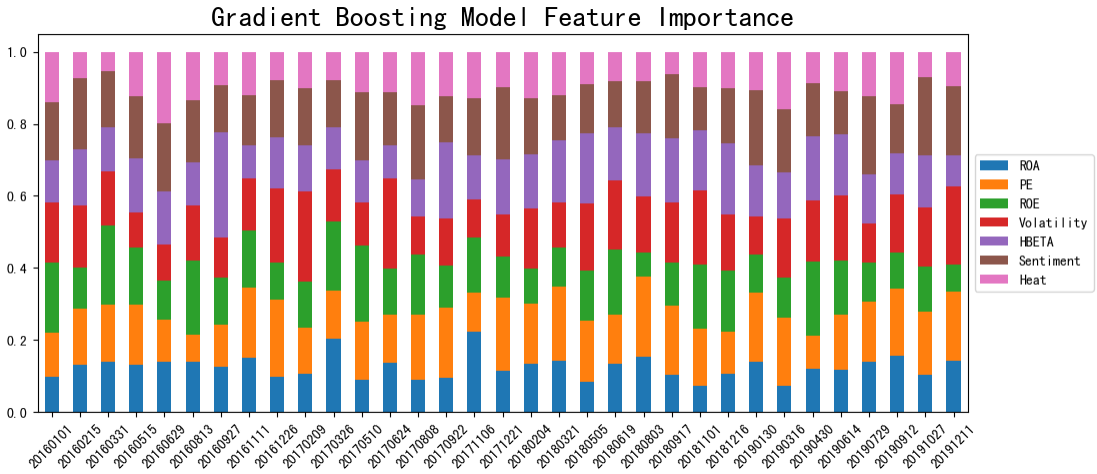


Figure 4: Factor weight

In this figure, we can find out that nearly all factors have the same weight,

emotion factor does not weight higher than other factors.

Conclusion

I also try the basic machine learning methods without emotion factor, and the results shows that the average annualized return is about 5%, which indicates emotion factor did increase the annuity return. At first, I want to compared these three methods with each other, and find the best one, and analysis the reasons, however, randomness was found in the back test results. In different back-testing period, we will have different back-testing results.

5. References

[1] Asness, Clifford S., Andrea Frazzini, and Lasse Heje Pedersen 2019 Quality Minus Junk. Review of Accounting Studies 24(1): 34–112.

[2] Baker, Malcolm, and Jeffrey Wurgler 2007 Investor Sentiment in the Stock Market. Journal of Economic Perspectives 21(2): 129–152.

[3] Brown, Gregory W., and Michael T. Cliff 2004 Investor Sentiment and the Near-Term Stock Market. Journal of Empirical Finance 1(11): 1–27.

[4] Campbell, John Y., and Robert J. Shiller 1988 The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors. The Review of Financial Studies 1(3): 195–228.

[5] Chen, Tianqi, and Carlos Guestrin 2016 XGBoost: A Scalable Tree Boosting System.

[6] Khaidem, Luckyson, Snehanshu Saha, and Sudeepa Roy Dey 2016 Predicting the Direction of Stock Market Prices Using Random Forest. ArXiv:1605.00003 [Cs]. http://arxiv.org/abs/1605.00003, accessed January 8, 2020.

[7] López-Salido, David, Jeremy C. Stein, and Egon Zakrajšek 2017 Credit-Market Sentiment and the Business Cycle*. The Quarterly Journal of Economics 132(3): 1373–1426.

[8] Mittal, Anshul, and Arpit Goel N.d. Stock Prediction Using Twitter Sentiment Analysis. http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=D6FD9503DA221E6B57140BF2FEEDA0B9?doi=10.1.1.375.4517, accessed January 8, 2020.

[9] Odean, Terrance 1998 Are Investors Reluctant to Realize Their Losses? The Journal of Finance 53(5): 1775–1798.

[10] Otoo, Maria Ward 1999 Consumer Sentiment and the Stock Market. SSRN Scholarly Paper, ID 205028. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=205028, accessed January 8, 2020.

[11] Tetlock, Paul C. 2007 Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance 62(3): 1139–1168.

[12] Zhou, Feng, Qun Zhang, Didier Sornette, and Liu Jiang 2019 Cascading Logistic Regression onto Gradient Boosted Decision Trees for Forecasting and Trading Stock Indices. Applied Soft Computing 84: 105747.

6.Appendix

This is the final project of Big Data course. You could find the source could here:

https://github.com/mengmeng12/ml_strategy

Word cloud visualization example:

https://mengmeng12.github.io/blog/wc_animation.mp4

https://mengmeng12.github.io/blog/wc_animation1.mp4

Codes available here:

https://github.com/mengmeng12/ml_strategy/tree/master/notebooks