Machine Learning Engined Asset Pricing: Empirical Evidence from China Financial Market

Jiawen Yan[*]
School of Economics and Management
Tsinghua University
yanjw.18@sem.tsinghua.edu.cn

Current Version: November 10, 2019

Abstract

This paper examines the effectiveness of various machine learning models in explaining and predicting returns in the Chinese stock market. We construct 91 firm-level factor loadings, extract 9 sentiment factors from textual data, and merge with 8 macroeconomic indexes from 2000 to 2018. Through a comprehensive comparative analysis among nine machine learning models, Partial Least Square (PLS), Ridge Regression (Ridge) and Linear Support Vector Regression (LSVR) are identified as the top three most successful models resulting lowest prediction errors and highest out-of-sample prediction accuracy. Characteristics of importance analysis further indicate that book-to-market ratio (BM), earnings-to-price ratio (EP), and industry momentum (INDMOM) are the top 3 most dominant predictive loadings as they explain most portion of return variances. Further, all machine learning models unambiguously agree that sentiment factors derived from textual contents are important asset-pricing factors in China. The predicting power of sentiment exists in both the text of analyst reports and social media posts but is strongest in traditional news media. Last, portfolio risk and return analysis proves machine learning models can generate real economic gains by constructing long-short equal-weighted decile portfolio to get an average of 3.40%, 2.64% and 1.92% monthly returns and 0.85, 0.66 and 0.49 annualized Sharpe ratios for PLS, Ridge, and LSVR, respectively.

Key Words: Machine Learning; Big Data; Sentiment Analysis; Fin-tech; Return Prediction; China Financial Market

[*] Jiawen Yan is a dual-degree master student in Business Analysis at Tsinghua University and Columbia University

"FinTech is not only an enabler but the driving engine"

- Pierre Gramegna, Minister of Finance of Luxembourg

1 Introduction

Asset pricing is a pivotal research branch in the financial market. Asset pricing theories determine the required or expected rate of return at a given level of risk on an asset. Two prominent asset pricing models describe the relationship between risk and return: Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT). CAPM model was proposed by Sharpe (1964) which states that for rational investors the expected return of an asset is determined by risk-free rate and the product between risk premium and the asset's systematic risk β. The APT model was invented by Roll and Ross (1984) suggesting that the market should be arbitrage-free and there are multiple risk factors that need to be considered when deriving risk-adjusted α. Asset pricing theories have fundamental influence in portfolio management, especially for asset allocation and rebalancing. Further, the asset pricing model 'helps us make an educated guess as to the expected return on assets that has not yet been traded in the market place.' (Bodie et al., 2005) and the implied risk-adjusted returns are widely used to discount future cash flows when calculating the present value (Balvers et al., 1990).

So important of asset pricing that numerous researchers are devoted to refine asset pricing models to better match risks and returns. A battery of related factors are analyzed and are added to the multi-factor model. The Fama-French model has three factors: size factor (SMB) and style factor (HML) besides existing market risk premium (Fama and French, 1996). Continuing studies explore the effectiveness of more factors such as liquidity and market momentum (Fama and French, 2016). Gu et al. (2018) summarize 94 factors throughout academia, and the factor zoo of Feng et al. (2018) lists as many as 147 existing empirically effective tradable factors. Following Gu et al. (2018), I constructed 91 firm-level pricing characteristics except for 3 not applicatable factors in China for all listed firms from 2000 to 2018.

To discover additional explaining sources, I explore the effectiveness sentiment factors. Sentiment related factors are novel factors catching significant attentions recently. In conventional context, sentiment broadly refers to the aggregated sentiment indexes obtained from surveys, such as consumer sentiment index and purchasing manager index in macroeconomic. However, both the number and the scope of surveyed sentiment indexes are vastly limited because of demanding capital and labor input. Yet, these obstructions can be overcame in recent days indirectly: we are observing an avalanche of user-generated unstructured data, mainly in textual form, on the Internet, enabling researchers to capture a variety of bodies' sentiment and to pinpoint links with diverse research questions. For example, Li (2010b) opens the door of modern textual analysis in accounting by using Naive Bayes algorithm to extract sentiment from 10-k files and documenting the existence of information content in annual reports; Huang et al. (2014a) shift context to analyst and prove the information content of textual reports in additional to contemporaneous quantitative numbers; Piotroski et al. (2015) massively collect millions of news articles and document the relationship between traditional media sentiment and stock volatility in China stock market; and Wang et al. (2019) further find that traditional media may be biased due to political concerns reflected in their attenuated market response, yet social media from East Guba, an aggregated grass root information platform, can correct this bias and provide a complementing role. In all, sentiment factors are proved to be effective. Thus, I derive sentiment factors from textual content generated by professional analysts, traditional media reporters and grass root investors at firm-, industry- and market-level, respectively, resulting in an additional 9 pricing factors. After manually introducing interaction terms between pricing factors and macroeconomic factors, and controlling for industry dummies, 989 pricing factors are inputted to pricing models.

Despite the vast existing literature on asset pricing, major studies are based on ordinary least squared (OLS) research paradigm and try to capture the naive linear relationship between innovative factors with market returns. A good factor, as agreed by consensus, should not only be statistically significant and able to improve coefficient of determination, R^2, but also have satisfying economic significance. Yet, this method has two drawbacks: (1) it fails to catch nonlinear relationships between factors and returns; and (2) R^2 only represents in-sample fitness but does not give clue on out-of-sample forecast accuracy. To deal with these concerns, machine learning models are introduced to the finance universe. By utilizing sophisticated penalizers and/or nonlinearities, machine learning models often have both higher prediction accuracy and generalisability. In this paper, I compare the results of 9 classic models in machine learning using data collected China financial market. They are Ridge Regression, ElasticNet, Partial Least Square, Stochastic Gradient Descent, Gradient Boost, Linear Support Vector Machine, and one- to three-layers Neural Networks.

To properly evaluate the performance of machine learning models, I take a leaf from statisticians' book. Liu et al. (2011) and Bennett et al. (2013) both recommend to use the mean absolute error (MAE), mean squared error (MSE) and variance explained (VE) as model fitness and prediction accuracy measures. Li (2017) further theoretically suggests that the correlation coefficient, r, and coefficient of determination, R^2, only capture in-sample goodness-of-fit, leaving prediction accuracy unknown and further making them improper measures for machine learning models. Hence, I do not use r or R^2 but comprehensively compare all three recommended measures for all models in different portfolios at different time horizons. All three measures agree that Linear Support Vector Machine, Stochastic Gradient Descent, and Ridge Regression are the top three outstanding models. Linear Support Vector machine can explain 0.58% out-of-sample variances and achieves an MAE of 10.18% and 190.08%^{2} for monthly returns.

Further, I compare variables of importance for firm pricing factors, macroeconomic pricing factors, and sentiment pricing factors. Empirical results suggest that book-to-market ratio (BM), earning-to-price ratio (EP) and industry momentum (INDMOM) are the top three most important pricing factors, contributing to more than 30% of all variance explained jointly. As for macroeconomic factors, major machine learning models agree that both market-level book-to-market ratio (market-BM) and market-level earning-to-price ratio (market-EP) are the top two most informative macroeconomic factors, contributing to near 50% of variances explained by all macroeconomic factors. And sentiment pricing factors derived from traditional news articles, especially at market-level, has the highest pricing importance (giving nearly 50% of all variance explained by all sentiment factors) over all other sentiment sources.

Finally, portfolio analysis indicates that Partial Least Square (PLS), Ridge Regression (Ridge) and Linear Support Vector Regression (LSVR) are top three models with annualized Sharpe ratio exceeding or nearing to 0.50, agreeing with model fitness measures discussed before. An average monthly return of 3.4% for Partial Least Square (PLS) model, 2.64% for Ridge Regression (Ridge) and 1.92% for the Linear Support Vector Regression (LSVR) model can be achieved by longing the top 10% and shorting the bottom 10% of all stocks. Though the Sharpe ratios are lower than U.S.'s comparable findings, China's financial market is well-known for its high volatility[1]. And it is undeniable that machine learning models can by and large produce relative satisfying economic gains.

In all, this paper contributes to asset pricing literature primarily in three folds. First, I reproduce 91 firm-level pricing characteristics following prior literature using China financial market data. Major existing studies focus on developed U.S. financial market but neglect emerging economy bodies such as China. Empirical evidences suggest that firm-level pricing factors jointly exhibit strong explaining power but have distinguishing chracteristics of importance when compared to U.S. market. Those well-constructed factors will greatly facilitate future related studies.

Second, this paper makes another data contribution by documenting the effectiveness of sentiment factors in China financial market. Through massively collecting TB size textual content of professional analyst reports and grass root investors' posts from the Internet, and by adopting an established sentiment database on traditional news articles, I find that all three sentiment sources exhibit significant pricing loadings on market. And, methodologically, Naive Bayes algorithm is an efficient and accurate textual analysis algorithm for identifying sentiment. Specifically, by labeling a relative tiny percentage of full data (around 10,000 sentences in this paper), this algorithm can produce a model with a decent and robust 78.16% out-of-sample predicting accuracy, and can then be deployed in parallel labeling full data.

Last, this paper sheds light on related studies by indicating that machine learning models are exceedingly effective in the setting of China financial market. Penalized linear models such as PLS, Ridge, and LSVR, though not as sophisticated as neural networks, are the most successful machine learning models. Yet, neural networks, which have most rewarding performance in mature markets, tend to overfit and do not produce satisfying results in China market. Overall, we identify several effective machine learning models and document their increasingly important roles in China financial market.

[1] See more at a Financial Times report about China market volatility, *"FT Explainer: Why are Chinas stock markets so volatile?"* by Josh Noble published on July 2, 2015

The following sections expand as follows: Section II reviews related studies about asset pricing, sentiment analysis and applications of machine learning models; Section III describes the technical details of used machine learning models; Section IV gives the data and summary statistics; Section V conducts empirical analysis through model fitness comparisons, variable of importance justifications and portfolio risk and return analysis; and Section VI concludes this paper.

2 Literature Review

2.1 Literature about Asset Pricing

Asset pricing has always been one of the most frequently discussed topics in the finance. Asset pricing theories attempt to establish the relationship between personal preference and capital market to determine the equilibrium rate of return, and to adjust asset price according to different risk preferences (Sharpe, 1964).

A majority of existing asset pricing literature investigates the problem in the framework of simply linear models. Initially proposed by Sharpe (1964), the capital asset pricing model, or CAPM model, states that the required rate of return can be unfolded into three parts: risk-free rate of return, the market return premium and the idiosyncratic risk factor, *Beta*. Fama and French (1992) modernize the CAPM model and propose the FF three-factor model by further including two additional risk factors - company size (SMB) and book-to-market (HML) - into consideration. Subsequent multi-factor researches in Titman and Jegadeesh (1993), Fama and French (2015) examine the effectiveness of more risk factors such as return momentum and market liquidity. The study paradigm above mentioned studies largely incorporates (1) constructing data in panels and introducing lagged pricing factors, (2) testing capital pricing factor significance under simply linear model assumptions, (3) constructing investment portfolios based on sorted lagged interested factors and (4) discussing risk and return tradeoffs between different portfolios (Fama and French, 2008, Lewellen, 2014). A few researchers also use time-series model to conduct predictive analysis using macroeconomic indicators to make time-series predictions of portfolio returns (Rapach and Zhou, 2013, Koijen and Van Nieuwerburgh, 2011).

2.2 Literature about Sentiment Analysis

Sentiment analysis actually is not a novel product, though heated discussed and studied in recent decades. Scholars have been trying to optimize consumer sentiment scaling models (Didow, Jr. et al., 1983); to understand the economic impact of consumer sentiment to general market (Gaski and Etzel, 1986); and to link general economic sentiment to other social consequences (Blood and Phillips, 1995).

Note that in the recent boom of the digital era, people rely more on Internet to acquire information and to express feelings. The readiness of more available data makes sentiment analysis one step further. With the help of big data, sentiment exhibits a powerful role in politics, for instance, Andranik et al. (2010) classified 104,003 Twitter posts in Germany into 12 categories including positive/negative emotion, sadness, anxiety, anger, certainty achievement etc., and found that microblogs, though as short as 140 characters, is an effective indicator for reflecting public political opinion and for predicting political election results with overall error smaller than 1.65%. Similar results are also found in different political settings (Bermingham and Smeaton, 2011, Wang et al., 2012).

The benefits of sentiments are also widely discussed in equity markets. Shleifer and Summers (1990) provided the rationale of the effectiveness of sentiment by raising the concept of *noise traders*, who make decisions not only on firms' fundamentals but also on their faiths or beliefs. In recent studies, scholars are trying to quantifying sentiment from textual data. Whether sentiment derived from textual data can provide additional information is still undecided in academia. Theoretically, if textual qualitative information purely explains existing quantitative numbers, then it should not provide additional information to investors (Francis and Soffer, 2006). Yet, empirical results suggest that investors will largely respond to several textual characteristics including textual sentiment, readability and confidence. Sentiment of annual reports (10-K) can be significant for forecasting future earnings (Li, 2010a), for asset pricing (Li, 2010a, Feldman et al., 2010, Davis et al., 2012, Jiang et al., 2019), for predicting analyst behaviors (Kothari et al., 2009), for forecasting cost of capital (Kothari et al., 2009) and for assessing firms' legal risks (Rogers et al., 2011). Sentiments of analyst reports are beneficial to investors as they provide additional information content besides existing quantitative forecasts such as EPS, recommendation and target price (Huang et al., 2014b). Besides, there is also strong evidence showing that direct sentiment of investors can be used to predict abnormal stock returns (Kumar, 2010, Hao, 2007, Frazzini and Lamont, 2008). In all, sentiment is playing an increasingly pivotal role in the financial market and can be helpful for asset pricing.

Needless to say the importance of properly measuring textual sentiment. This is typically done through natural language processing (NLP). Though many meaning strides have been made in general NLP tasks in recent years Mikolov et al. (2013), Pennington et al. (2014), sentiment analysis is still an uneasy task in finance. There are broadly two streams of literature about sentiment extraction. The first stream of studies uses the dictionary, a.k.a, keyword, method to measure the textual sentiment following Loughran and Mcdonald (2016). Though the dictionary method is easy to implement for academia people in finance and accouting, it is also discernable that it comes with three obvious drawbacks: 1) dictionary method relies on a pre-defined list of words, which can be subjective and lack of universality for uncommon/infrequent or domain-specific words; 2) dictionary method ignores the relationship between words - i.e. bigrams (two consecutive words) and trigrams (three consecutive words)[2], which can be very different when separated into multiple words; 3) time complexity of implementing dictionary method is relatively high[3]. In recent years, scholars have been trying to machine learning models to

[2] Such as Face Value as bi-gram and Earnings Per Share as tri-gram

[3] A typical implementation of dictionary method is done through iteratively counting the occurrence of

overcome those discussed problems and I will discuss this in detail in the next section.

2.3 Literature about Machine Learning

Machine learning and artificial intelligence are revolutionary products and they have been exhibiting exciting power and potential in real-world applications. Artificial intelligence upgrades and revolutionizes almost all products and services including communications, search engines, travel agents, health and healthcare, education, daily consumption, etc. As suggested in a report from the Wudaokou Global Financial Forum in 2019, the future 10 technical modules for the development of artificial intelligence are machine vision, speech and sound recognition, natural language processing, search engines, information processing and knowledge extraction, predictive analysis, planning and search agents, speech generation, image generation, manipulation and control, navigation and movement. Three key developments make these improvements possible. First, boosted computing power and storage space are becoming available in recent decades; Second, more well-annotated dataset with million of observations are emerging nowadays; Third, continuous developments of ML & AI algorithms and models are made by countless industrial experts and practitioners.

In natural language processing (NLP), the emergence and improvement of word vectors flip the page of natural language analysis. Mikolov et al. (2013) first proposed the concept of word vector and Word2Vec model with skip-gram and CBow [4]. The idea of word vector is to represent each word in the a high-dimensional vector space to reduce dimensionality and to avoid the common pitfall of dimension explosion faced by one-hot [5] representation. Pennington et al. (2014) from Stanford refined the Word2Vec algorithm and proposed the GloVe unsupervised learning vector model, which both greatly reduced training time and increased word vector representation accuracy by combining the idea of tf-idf and skip-gram. Subsequent downstream natural language tasks are often based on pre-trained word vectors on common corpuses, and fine-tuning vectors for more specific tasks. This gradually becomes industrial standard for various downstream tasks such as machine translation and human-machine dialogue (Mikolov, 2016).

each word of a given list in a document. To illustrate, suppose that a positive word list contains 200 words and a given document contains 100 sentences. To determine the sentiment level of each sentence, 100 × 200 = 20,000 times of comparisons are needed. The time complexity of this implementation is $O(n^2)$.

[4] In short, the Skip-gram model tries to predict surrounding words by using the center word; and CBow model tries to predict center word by using information of surrounding words. For example, for a given sentence "I have a red apple.", the CBow model tries to predict all surrounding words of "I", "have", "an" and "apple" given the center word "red"; while the Skip-gram model tries to predict the center word "red" given all surrounding words of "I", "have", "an" and "apple". Check Stanford CS224n for more about word vectors.

[5] One-hot representation uses binary indicators to denote for the appearance of each word, with 1 being appearance and 0 being non-appearance. For example, in a bag of word containing [I, you, we, have, has, a, an, apple, banana, pear], the one-hot representation of the sentence "I have an apple" is [1, 0, 0, 1, 0, 0, 1, 1, 0, 0]. With increased size of word dictionary (bag-of-words), the vector (for sentence) /matrix(for document) will quickly become very large and sparse, making further analysis difficult to perform.

In image recognition tasks, many creative deep networks were developed in recent years and the logic of going deep works really well. Russakovsky et al. (2015) collected millions of manually labeled images with category names and launched the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) competition in 2010, witnessing the booming of computer vision. Krizhevsky et al. (2017) reduced the error rate of ILSVRC tasks from previous 26.2% to 15.3% by proposing the visionary AlexNet model with multiple convolutional, pooled and fully connected layers and more than 6 million variables and 65,000 neurons. Since then, neural networks, especially sophisticated deep networks (such as recurrent Neural Network (RNN) and the Convolutional Neural Network (CNN) and their variants), are becoming accepted for their outstanding results and are being wildly used for a number of image and textual related tasks (Sagonas et al., 2015, Vernon et al., 2014). So successful for convolution and recurrent neural network, many financial practitioners are also eager to find out whether we can get good results using these novel models.

Machine learning and artificial intelligence technologies also demonstrate promising potential in capital market. They can facilitate the development of capital markets, improve the efficiency of information dissemination, and promote competition among institutions. It is also widely accepted that they will eventually subvert and reform the whole capital market. In recent accounting and finance literature, researchers are also trying to understand the effectiveness and implications of machine learning on capital market behaviors. One stream of studies approaches by retaining traditional linear paradigm and use machine learning model in part to find significant explaining variables. Major literature in this strand starts from traditional dismissed textual information and discusses the informativeness of textual information incremental to contemporaneous quantitative numbers. For example, Li (2010a) introduces the naive Bayesian machine learning method to the accounting field and applies it to textual content of annual report (10-K) of listed companies. Empirical evidence suggests that tone in management discussion and analysis (MD&A) section can provide investors with additional information besides existing quantitative fundamentals numbers. Huang et al. (2014b) employ Naive Bayesian machine learning approach to evaluate the sentiment level of text in U.S. S&P 500 analyst reports; through large sample study, they document that analysts' textual sentiments are correlated not only with short-term cumulative abnormal returns as well as with long-term company growth. Under the help of Latent Dirichlet Allocation (LDA) unsupervised machine learning model, Huang et al. (2014b) cluster and compare textual information in analyst reports and conference calls, further documenting analysts' information discovery role and information reinterpretation role.

Another strand of studies tries to tackle concerning questions directly with machine learning models. The underlying logic is that machine learning models can boost explaining power by introducing penalizing nonlinearities between cross-sectional explaining variables. Gu et al. (2018) first attempt to use machine learning model for asset pricing and they archive both satisfying out-of-sample forecast accuracy and high and stable long-short portfolio returns. Yet, they use firms' fundamental and technical indicators but ignore a significant part in the financial universe - sentiment. And It is still unknown whether those models will also have sounding performances in China financial market, an immature, volatile and policy-oriented market.

In all, machine learning and artificial intelligence have the potential of reshaping the traditional linear research paradigm and may provide us with unique understandings and deliberations of the capital market, especially in China's capital markets with few prior studies.

3 Theories and Models

3.1 Machine Learning Models

As accepted by computer science practitioners, machine learning model can achieve higher out-of-sample prediction accuracy and generalisability than the ordinary least square (OLS) model. Various ML models are designed and modified to accommodate different types of data in real-world settings. In this paper, I use nine machine learning models available in Scikit-Learn in Python[6]. Specifically, they are Ridge Regression (Ridge), Elastic Net (ENet), Partial Least Square (PLS), Stochastic Gradient Descent (SGD), Gradient Boost(GB), Linear Support Vector Regression(LSVR), and one- to three-layers Neural Networks(NN1-NN3). Note that I do not strictly follow the machine learning models used in Gu et al. (2018). The reasons are three-fold: (1) some machine learning models (such as PCR and GLM) are not directly available in Scikit-Learn Package and require both sophisticated programming skills and demanding time to implement those models efficiently; (2) a number of models (such as LSVR and SGD) are not included in the Gu et al. (2018) but they have been widely used and demonstrating exciting results in tasks such as image classification and sentiment analysis. Thus I add those models and further compare fitting results among the model zoo; (3) For multilayer perceptron model, a.k.a the neural network model, I use one- to three-layers on neural networks instead of one- to five-layers of networks in Gu et al. (2018) since anecdotal evidence suggests that the additional benefit of going deep is extremely limited for fully-connected network but requires significant longer training time as the number of combinations of model coefficients increase geometrically[7]. In the following sections, I will try to walk through each model and discuss the unique features/advantages of each vanilla machine learning model in a plain and intuitive manner.

3.1.1 Ridge Regression Model

Ridge Regression is a linear model similar to ordinary least square (OLS) but departs from it by penalizing model coefficients. Ridge model minimizes the objective function of:

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 + \alpha||\boldsymbol{w}||^2 \tag{1}$$

[6] See more about Scikit-Learn at https://scikit-learn.org.

[7] In fact, 4 and 5 layers neural networks are also fitted, yet the empirical results do not exhibit significant advantages comparing with shallow networks reported in this paper. Results are available upon request.

Note the first part $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2$ is just the squared sum of errors or the residuals in the OLS; the added latter part $\alpha||\boldsymbol{w}||^2$ is squared ($l2$) coefficients sizes, with α being the preset regularization intensity. Generally, the higher the α, the higher the penalization intensity and the smaller variation of model coefficients. This penalization intensity increases quadratically. By introducing the latter penalty term, Ridge model avoids very large coefficients caused by outliers, subsequently improving the conditioning of problem and reducing the variance of estimates.

3.1.2 ElasticNet Model

ElasticNet is also a variant of OLS linear model by combining both $l1$ and $l2$ regularization terms. In short, ElasticNet Model tries to minimize the objective of

$$\frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2}{2N} + l1\ ratio \times \alpha||\boldsymbol{w}|| + (1 - l1\ ratio) \times \frac{\alpha||\boldsymbol{w}||^2}{2} \tag{2}$$

Intuitively, the model tries to minimize both $l1$ and $l2$ regularizers jointly. As the pure $l1$ regularizer may be too weak and the pure $l2$ regularizer may be too harsh, the combination of the two tries to achieve a balance between two regularizers. Accordingly, there are two hyper parameters in this model: α and $l1\ ratio$. *Alpha* controls for the overall penalizing strength of the two regularizers and the mixing parameter $l1\ ratio$, sitting between 0 and 1, controls the weight between the two regularizers. The higher the α, the greater the penalty for model coefficients; while the higher the $l1\ ratio$, the more weight of penalty from $\alpha||\boldsymbol{w}||$ instead of $\alpha||\boldsymbol{w}||^2$. It is also interesting to note that ElasticNet model becomes Lasso model when $l1\ ratio$ is set to 1, in which the $l2$ regularization part is eliminated; and becomes Ridge model when $l1\ ratio$ is set to 0, in which the $l1$ regularization part is eliminated.

3.1.3 PLS Model

PLS is an enhanced and advanced model of PCR. The PCR, Principal Components Regression, regresses the interest dependent y on PCA scores: $\boldsymbol{X}^+ = \boldsymbol{P}_k(\boldsymbol{T}_k^T\boldsymbol{T}_k)^{-1}\boldsymbol{T}_k^T$. Yet, PCR comes with a major problem: major components are descriptive but not predictive, simply describing the variance of X, making the model easily stuck in local optima. PLS tries to address this issue by also considering properties of predicted y. PLS actually is a combination between PCR and MLR - the PCR captures the greatest variance of X while the MLR maximize the correlation between X and y - regressing y on $\boldsymbol{X}^+ = \boldsymbol{W}_k(\boldsymbol{P}_k\boldsymbol{W}_k)^{-1}(\boldsymbol{T}_k^T\boldsymbol{T}_k)^{-1}\boldsymbol{T}_k^T$. The weight $\boldsymbol{W}$ is added to maintain orthogonal scores. For each component k, this algorithm finds weights u and v that try to maximize the objective of: $corr(Xk\ u, Yk\ v) \times std(Xk\ u)\ std(Yk\ u)$ given $||u|| = 1$.

3.1.4 SGD Model

SGD regression is a linear model fitted by minimizing the loss with Stochastic Gradient Descent algorithm, which update model parameters along the way with a given learning rate. The loss function and penalty are customizable among a battery of choices. SGD optimizing algorithm employs the idea of batch training, which does not require holding all data into RAM, and it converges faster than as it performs parameter updates more frequently. Generally, stochastic descent algorithm has some random behaviors and those noise introduced during updating avoids the pitfall of jammed in local optima.

3.1.5 Gradient Boosting Model

The Gradient Boosting Model is an ensemble, or addictive, model that builds in a forward stage-wise manner. Given the number of estimators, the number of boosting stages to perform, a regression tree is fitted on the negative gradient of the given loss function at each stage. This multi-stage boosting idea takes advantage of ensemble, thus the joined model is typically robust to the overfitting problem, resulting in better performance than a single non-addictive model.

3.1.6 Linear Support Vector Regression Model

The Linear Support Vector Model tries to find a linear function $f(x)$ that a required accuracy ϵ is satisfied for every data point while minimizes the model complexity, w. One major distinction between SVR and linear regression is that the sum squared error term is minimized as best fit for the OLS model while a certain amount of error ϵ is allowed and does not account for error in SVR.

$$\begin{aligned} min \ \ & \frac{1}{2}||w||^2 \\ s.t. \ & ||y_i - Wx_i - b|| \leqslant \epsilon \end{aligned} \tag{3}$$

Support vector machines have a number of significant advantages: they are effective in high dimensional spaces and are also memory efficient by using a subset of training points in the decision function. Though many *kernel* functions are available for decision functions, I choose the linear kernel since it has the lowest model complexity and can yield comparable results against other kernels. The hyperparameter for this model is C, the penalty coefficient of the error term.

3.1.7 Neural Network

The Neural Network model, or Multi-layer Perceptron model (MLP), is the most well-known supervised machine learning model and is the vanilla model for many machine learning tasks. A typically neural network has three layers - input layer, hidden layers and output layer. The number of input layer and output layer is both one, but the number of hidden layers can be greater than one. The number of hidden layers is commonly referred to the depth of a neural network. We often think a model is a shallow network if it has two or less hidden layers and is a deep network if it has more than three hidden layers. In the vanilla model, parameters in the hidden layers are initialized randomly[8] and are updated under *backpropagation*. For each node, the input value is transformed by a preset activation function[9]. The activation function for each layer is customizable but it is important to keep at least one layer's activation function being non-linear since multiple linear activation functions can be transformed and combined into one, resulting in a pure linear function[10].

3.2 Model Fitness Measures

To assess the predictive accuracy of machine learning models, I compare the difference between the actual realized values and the predicted values. In statistics, numerous model fitness measures are used. The most common ones encompass: mean absolute error (MAE), mean square error (MSE), relative MAE (RMAE), root mean squared error (RMSE), variance explained (VE) etc., as summarized in Li (2017). The pure r, the correlation coefficient, and R^2, the coefficient of determination should *not* be used to assessing model prediction accuracy. The r and R^2, measuring the error between actual y and fitted value $\hat{y}$, only capture the goodness-of-fit between in sample y and x, making them improper and misleading for evaluating model prediction accuracy (Li, 2017). MAE and RMSE the most commonly used and recommended measures in Liu et al. (2011), Bennett et al. (2013). Since RMSE is simply the squared root of MSE, they are basically the same.

Though limitations are raised for both of above-mentioned measures, it is hard to deny that they are still easy-to-compute, valid and robust measures in this setting. This is guaranteed by data generating process of training and testing data, which are computed from the same database and are in the same unit/scale. As indicated in Equation 4, MAE is calculated as the sum of absolute value of the difference between actual and predicted values in out-of-sample (oos) set scaled by sample size N. Similarly, MSE is calculated as the sum of squared value of the difference between actual and predicted values in out-of-sample (oos) set scaled by sample size N, as exhibited in Equation 5.

$$MAE_{oos} = \frac{\sum_{oos} ||\hat{\beta}_t x_{i,t+1} - y_{i,t+1}||}{N} \tag{4}$$

$$MSE_{oos} = \frac{\sum_{oos} (\hat{\beta}_t x_{i,t+1} - y_{i,t})^2}{N} \tag{5}$$

[8] Other initialization approaches, such as, truncated normal, embedded vectors, etc. are available to avoid gradient exploding or diminishing problem.

[9] For example, for the ReLU (Rectangular Linear Unit) activation function, the output is 0 if input the x is smaller than 0 and is x is input x is greater or equal to 0.

[10] To illustrate this point, if the first activation function is $y = 0.5x+1$ and the second activation function is $z = 0.3y+2$, we may directly plug the first equation into the first equation, combining the two activations into $z = 0.15x + 2.3$.

Since both measures evaluate forecast errors but not accuracy (i.e., we know the how much error the models may produce, but we do not know the exact preciseness/accuracy of the models), I add a third model fitness measure - variance explained (VE) - to address this issue and further compare model prediction accuracy comprehensively. VE has a number of distinguishing excellent properties: it is a unit-free fitness measure, unifies several previous studied error measures, and measures both prediction error and prediction accuracy (Li, 2017). In Equation 6, VE is calculated as one minus the fraction between sum of squares explained, SSE, and sum of squares total, SST. A minor modification is made to SST by not demeaning the actual y by $\bar{y}$ because of reasonings discussed by Gu et al. (2018) for capital market data.

$$VE_{oos}^2 = 1 - \frac{\sum_{oos} (\hat{\beta}_t x_{i,t+1} - y_{i,t+1})^2}{\sum_{oos} y_{i,t+1}^2} \tag{6}$$

4 Data and Descriptive Statistics

The data used in this paper comes from three parts: (1) firm level technical and fundamental data; (2) macroeconomic data; and (3) sentiment data.

4.1 Technical and Fundamental Factors

Following literature listed in Gu et al. (2018), I build a large collection of stock-level predictive characteristics using China capital market data. Raw fundamental and technical data are downloaded from CSMAR database and the sample ranges from 2000m1 to 2018m12. The number of pricing factors I derived is 91 instead of 94 in Gu et al. (2018) due to the incompleteness of Chinese data[11]. Among the 91 firm-level factor loadings computed for empirical analysis from 2000 to 2018, 58 update annually, 13 update quarterly and 20 update monthly. I do not include samples prior to 2000 and the reasons are three-folds: first, China's capital market is still in its emerging stage and the number of observations with complete data is extremely small, bringing significant obstacles for the empirical analysis (Luo et al., 2018); second, most machine learning models work by learning and inferring on large data samples using statistical probabilities knowledge. The limited number of observations typically are not sufficient to train or tune the machine learning parameters, leading to the notorious $overfitting$ problem; third, major sentiment factors, especially that derived from online social media, do not come well-established until 2000 after the popularization of the Internet and mobile phones. Thus, data prior to 2000 is not only sporadic but also in poor quality, making it difficult, if not impossible, to use for empirical analysis. For the reasons justified above, I exclude data before 2000 from full sample. Refer to Appendix 6 for the original literature of 94 fundamental factors, the updating frequency and the corresponding CSMAR database; as well as the detailed

[11] In Gu et al. (2018), 94 firm-level factors are employed for asset pricing. Yet, three out of them (CONVIND, SECURED, SECUREDIND) are not derivable using financial data in China market.

variable meanings and respective CSMAR database files and fields for each firm-level pricing factor.

4.2 Macroeconomic Factors

Macroeconomic factors are included to adjust for macroeconomic variations at market-level. Following Welch and Goyal (2008) and Gu et al. (2018), I construct 8 macroeconomic factors to predict the stock price. Specifically, I use (1) DP, the dividend-to-price ratio; (2) EP, the earning-to-price ratio; (3) BM, the book-to-market ratio; (4) NTIS, the net equity expansion; (5) TBL, the treasury bill rate; (6) TMS, the term spread; (7) DFY, the default spread; and (8) SVAR, the stock variance. Each of them is derived using macro data available in CSMAR database.

4.3 Sentiment Factors

4.3.1 Data Source of Sentiment

So important of sentiment for behavior finance that I consider three categories of sentiment in this paper: sentiment of professional equity market sell-side analysts, sentiment of media and sentiment of public investors. Figure 1 provides a screenshot of a sample analyst report in China. Sentiment of analyst reports have been documented to be effective in capital market. By collecting over 363 thousands analyst reports of S&P500 firms, Huang et al. (2014b) document that professional analysts' tone derived from their textual reports can provide additional information content to investors besides contemporaneous quantitative numbers. Similarly, I design Python scripts to collect over 1.2 million China sell-side analyst reports from Tencent Finance and Eastmoney.com, two major public financial information providers in China from 2005 to 2018. Among all the analyst reports collected, 0.37 million is firm-level reports; 0.43 million is industry-level reports and the remaining 0.4 million is market-level reports. Each report is extremely detailed: HTML formatted webpage records structured information including publishing date, writing date, analyst name, affiliated brokerage, stock code, report title and textual content. I take the average of all analyst reports' sentiments for the same year, quarter and firm (industry, market) to derive the firm (industry, market) level analyst report sentiment factor, AR. par

For media sentiment, I use the sentiment of traditional news as a proxy. Figure 2 provides a screenshot of a financial news article covered by *Securities Times of China*, a government-owned media publishing news pertinent to public listed firms. Comprehensive China textual news sentiment dataset is available at *DataGo.com.hk*, which provides news sentiment covering more than 1000 traditional media, over 15 million news articles, from 1995 to 2018. Each news article is labeled with an adjusted sentiment index between [-1, 1] and I take the average of all news articles' sentiments for the same year, quarter and firm (industry, market) to derive the firm (industry, market) level traditional media level sentiment factor, TM.

Next, I use the sentiment of social media to measure public investors' sentiment. As stated in Bollen et al. (2011), sentiment on online platforms can be treated as proxy general investors' sentiment. To construct data, I design a parallel web crawler to collect 580 million investor discussions at Eastmoney.com, a Chinese social media platform allowing investors to post blogs and discuss news of public listed firms. Figure 3 provides a screenshot of a blog posted discussing future business and stock price outlook by a grass-root investor. I take the average of all investor posts' sentiments on social media for the same year, quarter and firm (industry, market) to derive the firm (industry, market) level social media sentiment factor, SM.
par

[Insert Figure 1, 2 & 3 about here]

4.3.2 Measurement of Sentiment

To properly extract sentiment from the above mentioned raw textual data, I employ the Naive Bayes machine learning method following Huang et al. (2014b). Specifically, there are three steps: (1) I firstly randomly select 10,000 sentences from the all textual information pool and manually label each sentence into one of the following three categories: positive, neutral or negative. After manually labeling, 2,386 of them are positive, 6,298 are neutral and 1,316 are negative; (2) Then I use the manually labeled sentences as golden truth and train a Naive Bayes[12] classifier[13] using Scikit-Learn and HanLP[14]; (3) Each sentence in the textual corpus is then labeled by the trained classifier and the sentiment of each document is calculated as the number of positive sentences minus the number of negative sentences adjusted by the length of text:

$$Document\ Sentiment = \frac{\#\ of\ Positive\ Sentences - \#\ of\ Negative\ Sentences}{\#\ of\ Total\ Sentences} \tag{7}$$

[12] In fact, Naive Bayes classifier has been proved to be one of the most successful algorithms in natural language processing and sentiment analysis (Lewis, 1998). I also compare the prediction accuracy of Naive Bayes classifier against 10 other popular machine learning algorithms such as Linear SVC, Logistic Regression, Random Forest, Stochastic Gradient Descent, Decision Trees, and One- to Four- Layers Neural Networks; multiple repeated tests agree that Naive Bayes algorithm outperforms other algorithms by achieving an average of 2% higher prediction accuracy.

[13] Here, ten-folds cross-validation accuracy of Naive Bayes classifier is 78.16% with a standard deviation less than 1%. 10-fold cross-validation proceeds as follows: randomly splitting data into ten parts, using 9 parts as training sample and the remaining part as testing sample. Since the testing sample is completely unseen during the training stage, the cross-validation accuracy reports the robust forecasting accuracy. Final accuracy is reported as the average of ten repeated tests.

[14] Unlike prior literature dealing with English content, Chinese words are not naturally separated by space. To address this problem, I use HanLP, a Python wrapped package to perform the required word-segmentation task. See more about HanLP at https://github.com/hankcs/HanLP

4.4 Data Preprocessing

Data preprocessing is often regarded as the most important step in machine learning as it makes the machine learning training phrase easier and results more reliable. To minimize the effect of irrelevant or redundant information and noisy or unreliable data, I perform the following data preprocessing steps. First, all continuous pricing factors are winsorzied at 1% and 99% level to minimize the effect of outliers and all missing data is replaced by contemporaneous industry median. Second, all continuous pricing factors regularized into range [-1, 1] following Gu et al. (2018) to reduced extreme effects on model parameters[15]. Third, considering the temporal order of data, I merge datasets using the following temporal logic: monthly updated factors are merged by lagging one month; quarterly updated factors are merged by lagging four months; annually updated factors are merged by lagging six months. This merging method avoids the common machine learning pitfall of introducing future data, making the output models more robust for out-of-sample prediction and forecasting[16].

As discussed in previous section, 91 pricing factors as well as 8 macro factors are derived. By further introducing analysts, news and social media sentiment at firm, industry and whole market-level, another 9 pricing factors are added. To allow machine learning models to learn from non-linearities between variables, (91+9) × 8 = 800 covariates are incorporated. I further add 81 China Securities Regulatory Commission Standardization Industry Classification (CIC) dummy variables as forecast indicators following Gu et al. (2018) to adjust for industry effects. All-inclusively, 91 (master pricing factors) + 9 (sentiment factors) + 8 (macroeconomic factors) + 800 (covariates between firm factors and macroeconomic factors) + 81 (CIC industry dummies) = 989 pricing factors are obtained for each company at each month. Altogether, 447,453 observations are included in the final sample. Table 9 reports detailed summary statistics of all factor loadings discussed in this section.

4.5 Model Training and Tuning

For machine learning models, three parts of data - training set, validation set and testing set - are needed. Training set is used to actually train the machine learning models and find statistical patterns among given data. Validation set is set apart for two purposes: first, to evaluate the model fitness during training stage and decide the time point at which training should stop; and second, to tune hyperparameters in machine learning models[17].

[15] As suggested in official documentation, this regularization approach is also recommended for major machine learning models in Scikit-Learn.

[16] For example, China public listed firms typically release annual report for year t before June 30 in year t+1. By employing the merging methodology mentioned above, observations prior to June 30 are merged with data released in annual report in year t-1; while post to June 30 are merged with data in year t. This ensures both precision and timeliness of factors inputted to models.

[17] Hyperparameter is a parameter of a prior distribution in Bayesian statistics. In machine learning, a hyperparameter is a parameter that is set before the learning process and the machine learning model

Hyperparameters are pre-defined parameters for machine learning models and they are typically not learned from training set but rather *tuned* through validation set. In this paper, I use the $grid-search$ method to iteratively search for the optimal hyperparameter from a given set for each model. The third part, testing data, is used to evaluate models' prediction ability and is completely unseen during training stage. Figure 4 depicts the above data rolling process vividly.

[Insert Figure 4 about here]

5 Empirical Analysis

5.1 Model Fitness

To evaluate the performance of trained machine learning models, I begin by comparing three goodness-of-fit measures discussed in previous section. Table 1 compares the model fitness for different models between different measures in different horizons. Panel A and B report monthly and annually mean absolute error, MAE, of nine employed machine learning models, with united of percent (%); Panel C and D report monthly and annually mean squared error, MSE, of nine employed machine learning models, with unit of percent squared ($\%^2$); and Panel E and F report monthly and annually the amount of variance explained, VE, a unit-less measure, of nine employed machine learning models in out-of-sample testing data, respectively. In each panel, the first row reports the model fitness of overall out-of-sample data; the second row of high group reports the model fitness of top 10% among predicted portfolio; while the last row of bottom group reports the model fitness of bottom 10% among predicted portfolio.

The results of the table have implications in three dimensions. First, comparing between different machine learning models, all model fitness measures agree that the Ridge Regression, the Stochastic Gradient Descent (SGD), and the Linear Support Vector Regression have less prediction error and higher prediction accuracy. LSVR has the most outstanding prediction outcome with monthly MAE being 10.18%, MSE being 190.08 $\%^2$ and VE being 0.63% in overall out-of-sample testing set. Further, all three measures, though measuring fitness in different dimensions, exhibit positive correlations and yield very similar empirical outcomes. It's surprising to find that 6 out of 9 models, including neural networks, have negative VEs, rendering those model largely ineffective explaining out-of-sample variances. Second, comparing between different time horizons of monthly and annually predicted returns, annually forecasts have lower prediction errors than monthly forecasts. Third, comparing forecast error between the overall testing sample with top and bottom deciles, forecasting errors for the top 10% decile are comparable

learn (or fit) the that given a the preset hyperparameter. For instance, Ridge Regression model minimize the objection function of $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 + \alpha||\boldsymbol{w}||^2$ and the α, a positive float, in the objective function is the hyperparameter of regularization strength, which improves the conditioning of the problem and reduces the variance s of estimates.

with full sample but are larger for the bottom 10% decile. Two reasons are plausible to justify this outcome: (1) there is no short-selling mechanism in China capital market and investors can only long positions, rendering the indexes more effective in upside than in downside; (2) bad events often come abruptly (such as the financial frauds of recent cases of KDX: 002450 and KMYY: 600518), making them extremely difficult, if not impossible, to predict from ordinary accounting factors computed from fundamental and technical data.

5.2 Variables of Importance Analysis

Despite satisfying results they may bring about, machine learning models are often been accused of operating as a black box for lacking interpretability. This section tries to tackle this problem by comparing characteristics of importance between different models. characteristics of importance evaluates the contribution of each pricing factor to prediction accuracy; and it is calculated the decrease of variance explained when a particular pricing factor is removed from the model. In the following sections, I will discuss variables of importance in three dimensions - firm-level pricing factors, macroeconomic variables and sentiment factors.

5.2.1 Variables of Importance of Pricing Factors

Figure 5 presents the variables of importance of 91 pricing factors among 9 employed machine learning models. The darker the color in each cell, the higher the importance of the pricing factor. Analyzing vertically regarding this two dimensional figure, pricing factors are sorted according to their relative importance from top to bottom and almost all models agree that top five most important pricing factors are book-to-market ratio (BM), earning-to-price ratio (EP), industry momentum (INDMOM), trading volume (DOLVOL) and accruals (ACC). Horizontal analysis captures the factor importance difference among nine utilized machine learning models. Ridge regression and Gradient Boost are rather distinctive models, leaning on R&D expenses and other factors that are not as weighted as in other models.

Figure 6 quantitatively presents the top 20 important pricing factors by each model. Variables of importance are normalized to sum to one for each model for better interpretation of their relative importance. BM, EP and INDMOM contribute to more than 30% of all variance explained.

5.2.2 Variables of Importance of Macro Variables

Table 2 and Figure 7 shows the importance of microeconomic forecasting factors. All models except for Gradient Boost (GB) agree that market-level book-to-market ratio (market-BM) and market-level earning-to-price ratio (market-EP) are the two most important forecasting loadings, contributing to near 50% of variances explained by all macroeconomic factors. Other factors bring about less pivotal predicting power. Gradient Boost method is a rather separated model that pins more importance on the market stock price variance (SVAR).

5.2.3 Variables of Importance of Sentiment Variables

Table 3 and Figure 8 reveal the importance of sentiment factors. It is not unsurprising that general market sentiment of news articles dominates other sentiment factors, giving nearly 50% of all variance explained by sentiment factors. Three perceivable patterns can be observed in Figure 8. First, sentiment pricing factors generated from news articles have higher importance than analysts' and public investors' sentiment at all stock-, industry-, and market-level. Second, market sentiment pricing factors account for much higher factor importance than both stock and industry level sentiment pricing factors. And third, different models put roughly similar emphasis on various sentiment factors. Partial least square (PLS), Gradient Boost (GB) and Linear support vector regression (LSVR) place slightly more emphasis on market-level sentiment at all analysts, news and social media sentiment, making them potentially more advantageous in forecasting out-of-sample stock returns.

5.3 Portfolio Risk and Return

Table 4 reports the performance of equal-weighted decile of machine learning portfolios. Instead of simply comparing for individual stock returns, monthly returns are sorted on model prediction results and are equally divided into ten investment portfolios from lowest to highest. Actual realized returns, standard deviations and annualized Sharpe ratios are computed accordingly for portfolio-level analysis.

Portfolio analysis is conducted out of three practical concerns. First, many investors make indirect investments through mutual funds and private funds who hold a battery of risky-assets. Portfolio analysis, thus, is more aligned with reality and investors' economic interests. Second, portfolio analysis also indirectly estimates model performance on a broader perspective as training and testing are based on stock-level observations. Third, portfolio analysis gives us a better insight between risks and returns of investment targets by allowing us to compare the relationship between volatility and returns in different portfolios.

As results in Table 4 indicate, Partial Least Square (PLS), Ridge Regression (Ridge) and Linear Support Vector Regression (LSVR) are top three models with annualized Sharpe ratio exceeding or nearing to 0.50. This result suggests that machine learning results are robust to different performance measures and all three above-mentioned models have both excellent individual stock predictability and outstanding portfolio performance. Average monthly return generated by long-short strategy is 3.40% for Partial Least Squared (PLS) model, 2.64% for Ridge Regression (Ridge) and 1.92% for the Linear Support Vector Regression (LSVR) model. Elastic Net (ENet) and Stochastic Gradient Boost (SGD) model both penalize coefficients too harsh that their predicted returns only varying between range (-0.2% and 0.2%). Neural network 1 - 3 (NN1, NN2, NN3) all encounter some sorts of overfitting problems. They all have excellent performance on training data but fail to translate the predicted good results into real economic gains. One-layer neural network model has an average predicted monthly return of 2.39% but only exhibit 0.32% return for the long-short portfolio. This is akin to high explaining power in training set but significantly loss explaining ability in testing, a.k.a., overfitting. Similar patterns are also found in two-layer and three-layer neural networks.

Comparing with results in Gu et al. (2018), there are several distinguishing features about performance of machine learning models in China financial market in China. First, market volatility is notoriously high, making machine learning model less effective. In U.S. market, ML predicted and realized standard deviation of monthly returns from hedging portfolios are typically less than 6%; while the number doubled with China's data. Second, the following result of highly volatile market returns is low realized Sharpe ratio. The machine learning models can often generate Sharpe ratios exceed 1.5 and even 2.0 from neural network models. Nevertheless, the highly volatile market not only makes neural networks ineffective but also drags down model Sharpe ratios. Third, the return volatility differences between deciles are significantly smaller when compared to U.S. results. Standard deviations of top and bottom portfolios return are 10% to 20% larger than middle-situated portfolios in the U.S. Yet the curvature is more flat in China with similar variations often less than 10%. This is potentially due to the daily return volatility limit in China capital market (up and down 10% limit). Fourth, the predicted returns and realized returns are similar in U.S. but are not so comparable in China even though they may produce acceptable results. This suggests that machine learning models are easier to suffer from overfitting problem in China since fewer data are available both cross-sectionally and time-seriesly. Also, this additionally implies that traditional machine learning models with slightly more penalization to plain OLS can produce satisfying results while neural networks with tons of parameters need to be fed with real big data to ensure them to work well.

6 Conclusions

In this paper, I examine the applications of various machine learning models in the setting of Chinese financial market. By using China market data from CSMAR, I construct 91 firm-level fundamental and 8 macroeconomic pricing loadings following previous literature. 9 additional sentiment factors are also constructed, covering tones of professional analysts from analysts reports, of news articles from traditional media and public investors from social media, at all stock, industry and market-level. Empirical evidence from 9 machine learning models of Ridge regression, Elastic Net, Partial Least Square, Stochastic Gradient Descent, Gradient Boost, Linear Support Vector Regression and one- to three-layers Neural Networks, suggests that LSVR, PLS and Ridge are the three best-performing machine learning models with all MAE, MSE and VE error measures being better than other models. Variables of importance analysis indicates that book-to-market ratio (BM), earning-to-price ratio (EP), industry momentum (INDMOM), trading volume (DOLVOL), and accruals (ACC) are top five most important firm-level pricing loadings; that market-level book-to-market ratio and market-level earning-to-price ratio are two most contributing macroeconomic factors; and that market-level sentiment indexes have significantly more predicting power than firm and industry level pricing loadings. Moreover, portfolio analysis suggests that PLS, Ridge and LSVR models can bring about 3.40%, 2.64% and 1.92% monthly returns through constructing a long-short investment portfolio. And the annualized Sharpe Ratios are 0.85, 0.66 and 0.49, accordingly.

Taken together, this paper proves the effectiveness of machine learning models in China financial market. Engined with fin-tech and novel machine learn models, we can revisit previous research questions in distinguishing perspectives and have a constructive understanding to the financial market.

References

Andranik, Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe, 2010, Predicting Elections with Twitter: What: What 140 Characters Reveal about Political Sentiment Andranik, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media Predicting* 178–185.

Balvers, Ronald J, Thomas F Cosimano, and Bill Mcdonald, 1990, American Finance Association Predicting Stock Returns in an Efficient Market Predicting Stock Returns in an Efficient Market, *Source: The Journal of Finance THE JOURNAL OF FINANCE * VOL. XLV* 45, 1109–1128.

Bennett, Neil D., Barry F.W. Croke, Giorgio Guariso, Joseph H.A. Guillaume, Serena H. Hamilton, Anthony J. Jakeman, Stefano Marsili-Libelli, Lachlan T.H. Newham, John P. Norton, Charles Perrin, Suzanne A. Pierce, Barbara Robson, Ralf Seppelt, Alexey A. Voinov, Brian D. Fath, and Vazken Andreassian, 2013, Characterising performance of environmental models, *Environmental Modelling and Software* .

Bermingham, Adam, and Alan F Smeaton, 2011, On Using Twitter to Monitor Political Sentiment and Predict Election Results, *Sentiment Analysis where AI meets Psychology (SAAIP) Workshop at the International Joint Conference for Natural Language Processing (IJCNLP)* 2–10.

Blood, DJ, and PCB Phillips, 1995, Recession Headline News, Consumer Sentiment, The State of the Economy and Presidential . . . .

Bodie, Z, A Kane, and A J Marcus, 2005, *Investments*, Irwin series in finance (McGraw-Hill).

Bollen, Johan, Huina Mao, and Xiaojun Zeng, 2011, Twitter mood predicts the stock market, *Journal of Computational Science* 2, 1–8.

Davis, Angela K., Jeremy M. Piger, and Lisa M. Sedor, 2012, Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language, *Contemporary Accounting Research* 29, 845–868.

Didow, Jr., Nicholas M., William D. Perreault, Jr., and Nicholas C. Williamson, 1983, A Cross-Sectional Optimal Scaling Analysis of the Index of Consumer Sentiment, *Journal of Consumer Research* 10, 339.

Fama, Eugene, and Kenneth R. French, 1992, The CrossSection of Expected Stock Returns, *The Journal of Finance* .

Fama, Eugene F, and Kenneth R French, 1996, Multifactor explanations of asset pricing anomalies-Journal of Finance, *Journal of Finance* 51, 55–84.

Fama, Eugene F, and Kenneth R French, 2008, Dissecting Anomalies, *The Journal of Finance* 63, 1653–1678.

Fama, Eugene F., and Kenneth R. French, 2015, Five-Factor Asset Pricing Model, *Journal of Financial Economics* 52.

Fama, Eugene F., and Kenneth R. French, 2016, Dissecting Anomalies with a Five-Factor Model, *Review of Financial Studies* .

Feldman, Ronen, Suresh Govindaraj, Joshua Livnat, and Benjamin Segal, 2010, Management's tone change, post earnings announcement drift and accruals, *Review of Accounting Studies* 15, 915–953.

Feng, Guanhao, Stefano Giglio, Dacheng Xiu, Alex Belloni, John Campbell, John Cochrane, Chris Hansen, Lars Hansen, Bryan Kelly, Stefan Nagel, and Chen Xue, 2018, Taming the Factor Zoo: a Test of New Factors, *Working Paper* .

Francis, Jennifer, and Leonard Soffer, 2006, The Relative Informativeness of Analysts' Stock Recommendations and Earnings Forecast Revisions, *Journal of Accounting Research* 35, 193.

Frazzini, Andrea, and Owen A. Lamont, 2008, Dumb money: Mutual fund flows and the cross-section of stock returns, *Journal of Financial Economics* 88, 299–322.

Gaski, John F., and Michael J. Etzel, 1986, The Index of Consumer Sentiment toward Marketing, *Journal of Marketing* 50, 71.

Gu, Shihao, Bryan T. Kelly, and Dacheng Xiu, 2018, Empirical Asset Pricing via Machine Learning, *Ssrn* .

Hao, Wang, 2007, Investor Sentiment in the Stock Market, *Ssrn* 21, 129–151.

Huang, Allen, Reuven Lehavy, Amy Zang, and Rong Zheng, 2014a, Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach, *Ssrn* .

Huang, Allen H., Amy Y. Zang, and Rong Zheng, 2014b, Evidence on the Information Content of Text in Analyst Reports, *The Accounting Review* 89, 2151–2180.

Jiang, Fuwei, Joshua Lee, Xiumin Martin, and Guofu Zhou, 2019, Manager sentiment and stock returns, *Journal of Financial Economics* 132, 126–149.

Koijen, Ralph S.J., and Stijn Van Nieuwerburgh, 2011, Predictability of Returns and Cash Flows, *Annual Review of Financial Economics* 3, 467–491.

Kothari, S. P., Xu Li, and James E Short, 2009, The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis, *The Accounting Review* 84, 1639–1670.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, 2017, ImageNet classification with deep convolutional neural networks, *Communications of the ACM* 60, 84–90.

Kumar, Alok, 2010, Self-selection and the forecasting abilities of female equity analysts, *Journal of Accounting Research* .

Lewellen, Jonathan, 2014, The Cross Section of Expected Stock Returns, *SSRN Electronic Journal* 47, 427–465.

Lewis, David D, 1998, Naive (Bayes) at forty: The independence assumption in information retrieval, number x, 4–15.

Li, Feng, 2010a, Textual Analysis of Corporate Disclosures: A Survey of the Literature.

Li, Feng, 2010b, The information content of forward- looking statements in corporate filings-A naïve bayesian machine learning approach, *Journal of Accounting Research* 48, 1049–1102.

Li, Jin, 2017, Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what?, *PLoS ONE* 12, 1–16.

Liu, Canran, Matt White, and Graeme Newell, 2011, Measuring and comparing the accuracy of species distribution models with presence-absence data, *Ecography* .

Loughran, Tim, and Bill Mcdonald, 2016, Textual Analysis in Accounting and Finance: A Survey, *Journal of Accounting Research* 54, 1187–1230.

Luo, Mei, Shuai Shao, and Frank Zhang, 2018, Does financial reporting above or below operating income matter to firms and investors? The case of investment income in China, *Review of Accounting Studies* 23, 1754–1790.

Mikolov, Armand Joulin; Edouard Grave; Piotr Bojanowski; Tomas, 2016, Bag of Tricks for Efficient Text Classification, *arXiv:1607.01759v3* .

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013, Distributed Representations of Words and Phrases and their Compositionality, *CrossRef Listing of Deleted DOIs* 1.

Pennington, Jeffrey, Richard Socher, and Christopher Manning, 2014, Glove: Global Vectors for Word Representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543 (Association for Computational Linguistics, Stroudsburg, PA, USA).

Piotroski, Joseph D., T. J. Wong, and Tianyu Zhang, 2015, Political incentives to suppress negative information: Evidence from Chinese listed firms, *Journal of Accounting Research* 53, 405–459.

Rapach, David, and Guofu Zhou, 2013, Forecasting Stock Returns, in *Handbook of Economic Forecasting*, volume 2, 328–383 (Elsevier B.V.).

Rogers, Jonathan L., Andrew Van Buskirk, and Sarah L C Zechman, 2011, Disclosure tone and shareholder litigation, *Accounting Review* 86, 2155–2183.

Roll, Richard, and Stephen A. Ross, 1984, The Arbitrage Pricing Theory Approach to Strategic Portfolio Planning, *Financial Analysts Journal* 40, 14–26.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, 2015, ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision* 115, 211–252.

Sagonas, Christos, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic, 2015, 300 Faces In-The-Wild Challenge: database and results, *Image and Vision Computing* .

Sharpe, W F, 1964, Capital Asset Prices:, *A Theory Of Market Equilibrium Under Conditions Of Risk. Journal Of Finance, 19, Pp* .

Shleifer, Andrei, and Lawrence H Summers, 1990, The Noise Trader Approach to Finance, *Journal of Economic Perspectives* 4, 19–33.

Titman, Sheridan, and Narisimhan Jegadeesh, 1993, Returns to Buying Winners and Selling Losers : Implications for Stock Market Efficiency, *The Journal of Finance* 48, 65–91.

Vernon, R. J. W., C. A. M. Sutherland, A. W. Young, and T. Hartley, 2014, Modeling first impressions from highly variable facial images, *Proceedings of the National Academy of Sciences* .

Wang, Eric, Tianyu Zhang, and TJ Wong, 2019, Do Chinese Social Media Delineate the Optimistic Bias of Traditional Media ?, *MIT Asia Conference* .

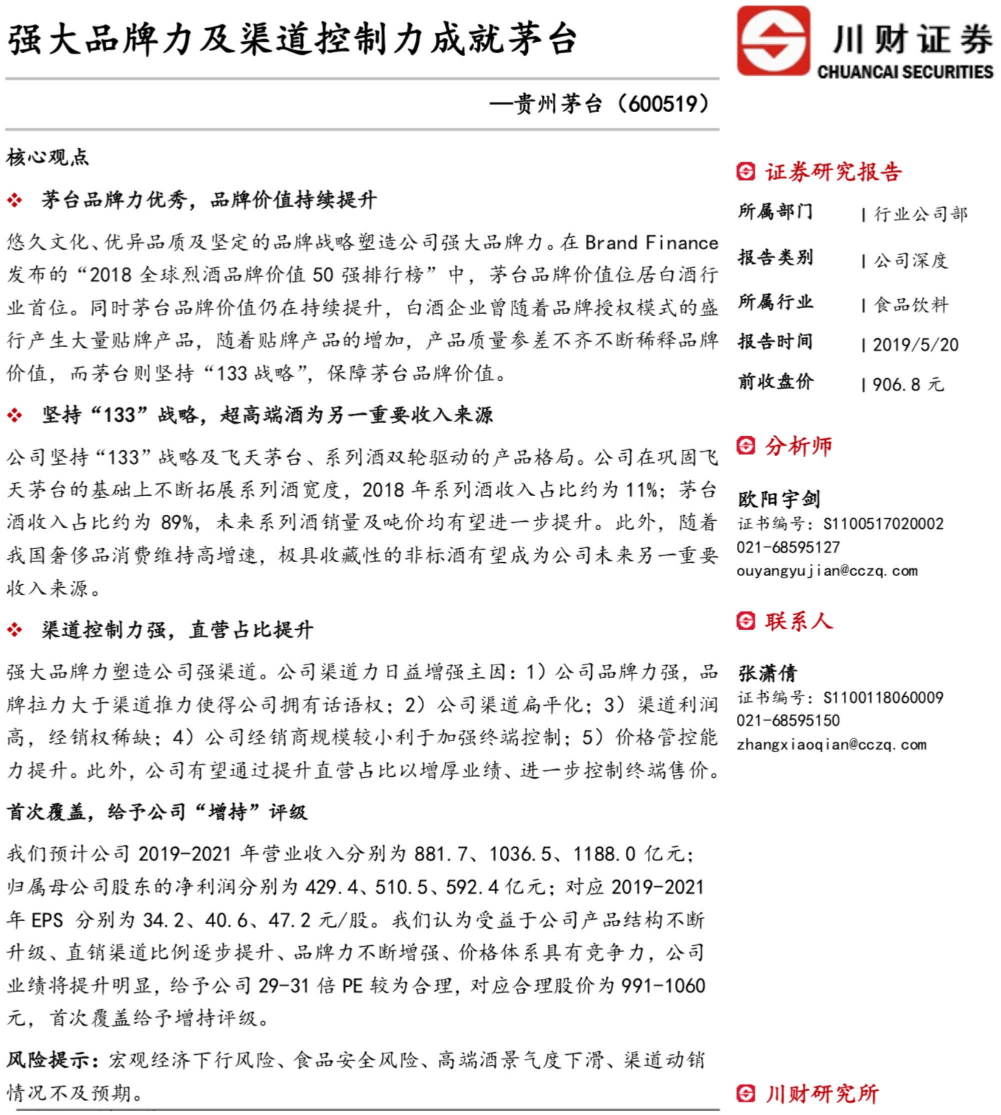
Wang, H., D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, 2012, A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* 115–120.

Welch, Ivo, and Amit Goyal, 2008, A Comprehensive Look at The Empirical Performance of Equity Premium Prediction, *Review of Financial Studies* 21, 1455–1508.

Appendix

1. Screenshot of Sentiment Sources

1.1 Screenshot of Analyst Reports



强大品牌力及渠道控制力成就茅台

—贵州茅台（600519）

川财证券
CHUANCAI SECURITIES

核心观点

❖ **茅台品牌力优秀，品牌价值持续提升**

悠久文化、优异品质及坚定的品牌战略塑造公司强大品牌力。在 Brand Finance 发布的“2018 全球烈酒品牌价值 50 强排行榜”中，茅台品牌价值位居白酒行业首位。同时茅台品牌价值仍在持续提升，白酒企业曾随着品牌授权模式的盛行产生大量贴牌产品，随着贴牌产品的增加，产品质量参差不齐不断稀释品牌价值，而茅台则坚持“133 战略”，保障茅台品牌价值。

❖ **坚持“133”战略，超高端酒为另一重要收入来源**

公司坚持“133”战略及飞天茅台、系列酒双轮驱动的产品格局。公司在巩固飞天茅台的基础上不断拓展系列酒宽度，2018 年系列酒收入占比约为 11%；茅台酒收入占比约为 89%，未来系列酒销量及吨价均有望进一步提升。此外，随着我国奢侈品消费维持高增速，极具收藏性的非标酒有望成为公司未来另一重要收入来源。

❖ **渠道控制力强，直营占比提升**

强大品牌力塑造公司强渠道。公司渠道力日益增强主因：1）公司品牌力强，品牌拉力大于渠道推力使得公司拥有话语权；2）公司渠道扁平化；3）渠道利润高，经销权稀缺；4）公司经销商规模较小利于加强终端控制；5）价格管控能力提升。此外，公司有望通过提升直营占比以增厚业绩、进一步控制终端售价。

首次覆盖，给予公司“增持”评级

我们预计公司 2019-2021 年营业收入分别为 881.7、1036.5、1188.0 亿元；归属母公司股东的净利润分别为 429.4、510.5、592.4 亿元；对应 2019-2021 年 EPS 分别为 34.2、40.6、47.2 元/股。我们认为受益于公司产品结构不断升级、直销渠道比例逐步提升、品牌力不断增强、价格体系具有竞争力，公司业绩将提升明显，给予公司 29-31 倍 PE 较为合理，对应合理股价为 991-1060 元，首次覆盖给予增持评级。

风险提示：宏观经济下行风险、食品安全风险、高端酒景气度下滑、渠道动销情况不及预期。

证券研究报告

所属部门	| 行业公司部
报告类别	| 公司深度
所属行业	| 食品饮料
报告时间	| 2019/5/20
前收盘价	| 906.8 元

分析师

欧阳宇剑
证书编号：S1100517020002
021-68595127
ouyangyujian@cczq.com

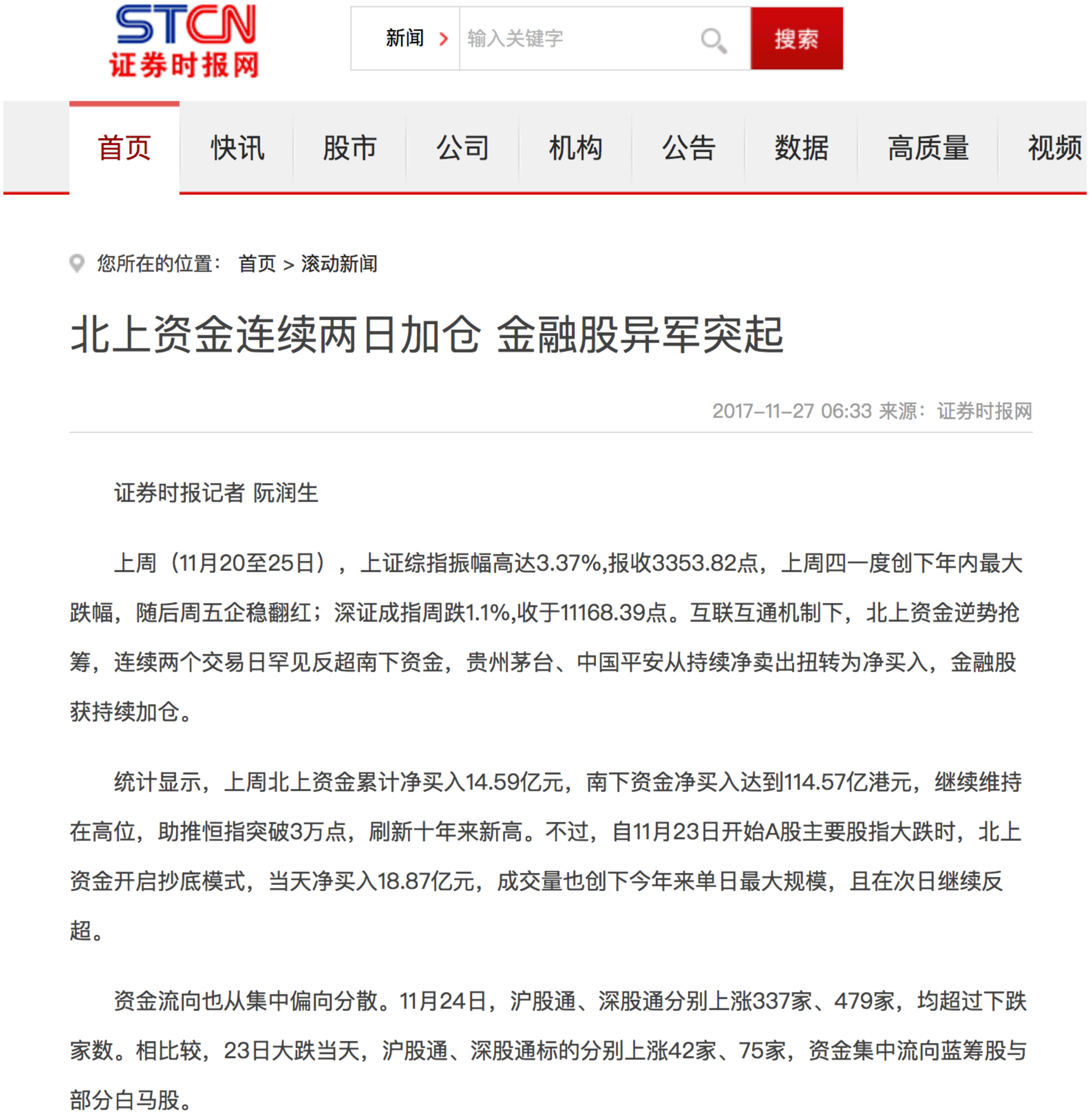
联系人

张潇倩
证书编号：S1100118060009
021-68595150
zhangxiaoqian@cczq.com

川财研究所

Figure 1: Screenshot of Analyst Reports

1.2 Screenshot of News Article



北上资金连续两日加仓 金融股异军突起

2017-11-27 06:33 来源：证券时报网

证券时报记者 阮润生

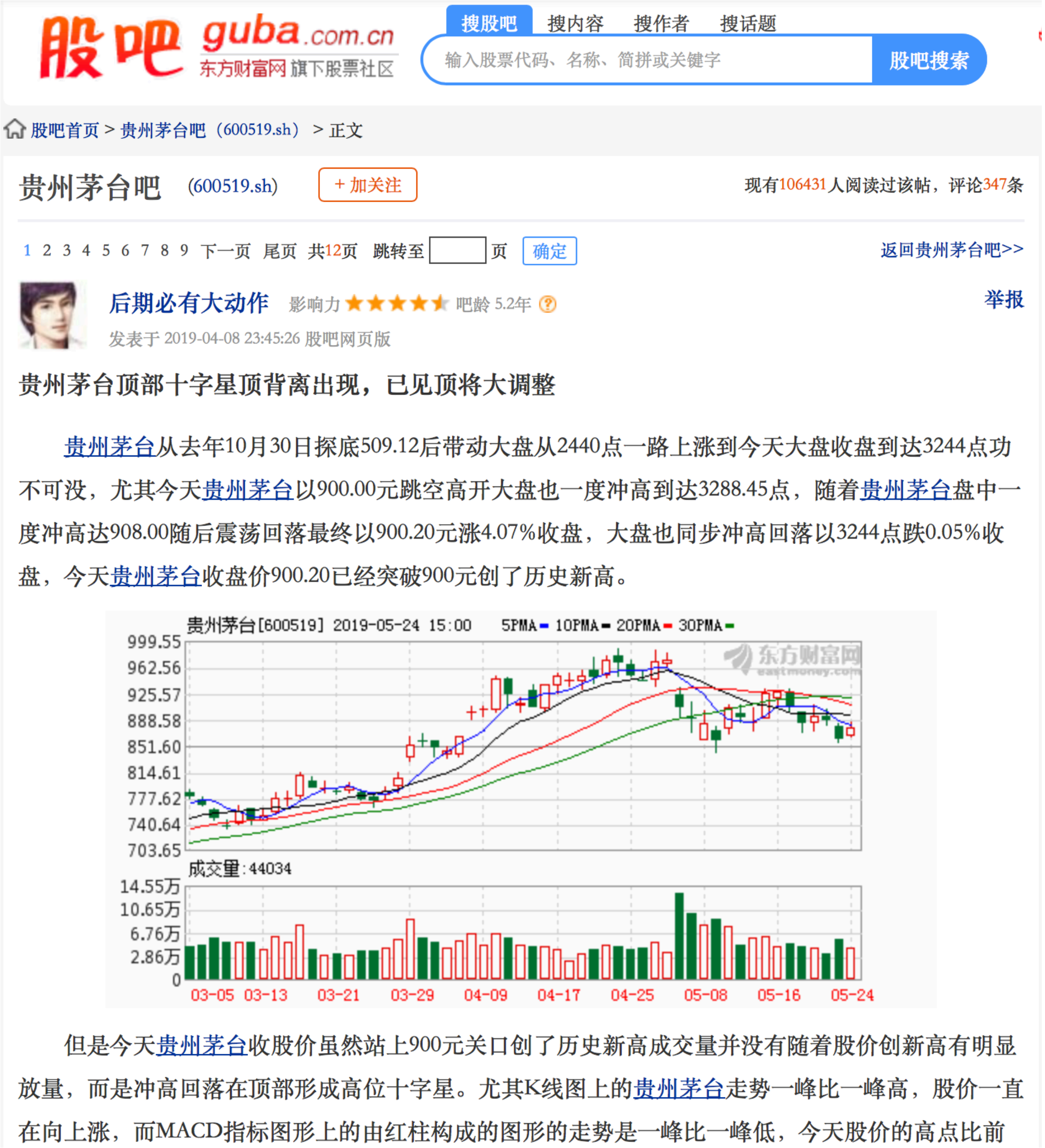
上周（11月20至25日），上证综指振幅高达3.37%,报收3353.82点，上周四一度创下年内最大跌幅，随后周五企稳翻红；深证成指周跌1.1%,收于11168.39点。互联互通机制下，北上资金逆势抢筹，连续两个交易日罕见反超南下资金，贵州茅台、中国平安从持续净卖出扭转为净买入，金融股获持续加仓。

统计显示，上周北上资金累计净买入14.59亿元，南下资金净买入达到114.57亿港元，继续维持在高位，助推恒指突破3万点，刷新十年来新高。不过，自11月23日开始A股主要股指大跌时，北上资金开启抄底模式，当天净买入18.87亿元，成交量也创下今年来单日最大规模，且在次日继续反超。

资金流向也从集中偏向分散。11月24日，沪股通、深股通分别上涨337家、479家，均超过下跌家数。相比较，23日大跌当天，沪股通、深股通标的分别上涨42家、75家，资金集中流向蓝筹股与部分白马股。

Figure 2: Screenshot of News Article

1.3 Screenshot of Social Media



股吧首页 > 贵州茅台吧（600519.sh） > 正文

贵州茅台吧 (600519.sh) + 加关注 现有106431人阅读过该帖，评论347条

1 2 3 4 5 6 7 8 9 下一页 尾页 共12页 跳转至 页 确定 返回贵州茅台吧>>

后期必有大动作 影响力 ★★★★★ 吧龄 5.2年 举报

发表于 2019-04-08 23:45:26 股吧网页版

贵州茅台顶部十字星顶背离出现，已见顶将大调整

贵州茅台从去年10月30日探底509.12后带动大盘从2440点一路上涨到今天大盘收盘到达3244点功不可没，尤其今天贵州茅台以900.00元跳空高开大盘也一度冲高到达3288.45点，随着贵州茅台盘中一度冲高达908.00随后震荡回落最终以900.20元涨4.07%收盘，大盘也同步冲高回落以3244点跌0.05%收盘，今天贵州茅台收盘价900.20已经突破900元创了历史新高。

但是今天贵州茅台收股价虽然站上900元关口创了历史新高成交量并没有随着股价创新高有明显放量，而是冲高回落在顶部形成高位十字星。尤其K线图上的贵州茅台走势一峰比一峰高，股价一直在向上涨，而MACD指标图形上的由红柱构成的图形的走势是一峰比一峰低，今天股价的高点比前

Figure 3: Screenshot of Social Media

2. Data Rolling Illustration

Training set, validation set and testing set are rolled in accordance with Figure 4 illustrated below. Beginning from 2000, I use every two years of data for training and one following year of data for validation and one next year of data for out-of-sample testing. Note that I roll validation set and testing set annually as well as training set, not putting *all* historical data available to training set as Gu et al. (2018). This comes with two deliberations: (1) China's capital market shows distinctive characteristics in each stage, adding data far from present may increase the noise and jeopardize the model accuracy; (2) Reduced training data increase both training and tuning speed, greatly reducing the computational power need.

For instance, standing at the end of 2003, I use 2000 and 2001's data to train various machine learning models and use 2002's data to tune hyperparameters for models. The trained models are then evaluated using 2003's data. Next, I roll all training, validation and testing set forward by one year and repeat the above processes. I refit the model annually following Gu et al. (2018). In fact, empirical evidence indicates that the additional model fitness is extremely limited if models are refitted monthly considering that sentiment factors that update monthly.

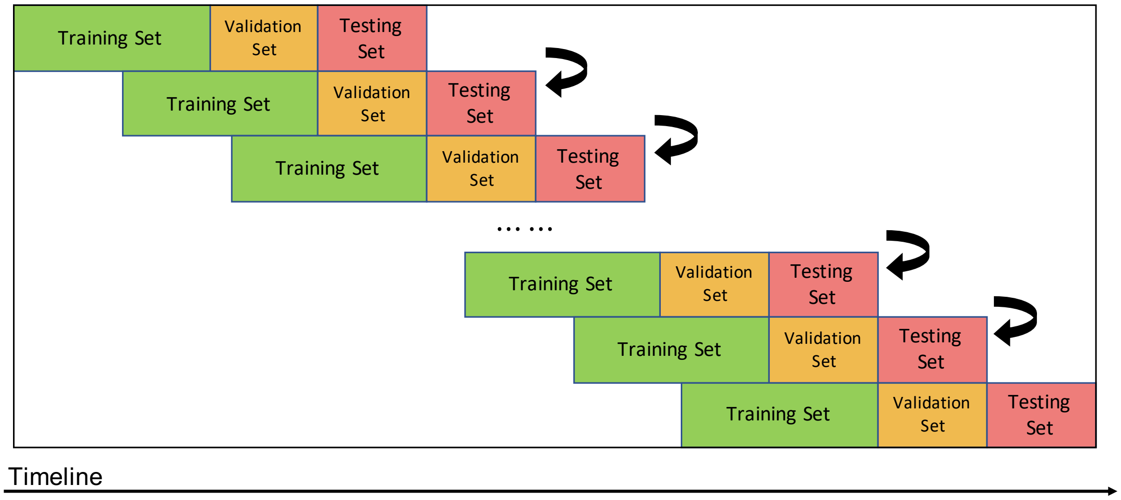


Figure 4: Training, Validation and Testing Data Rolling Illustration

3. Model Fitness between Different Measures

Note: Table 1 compares model fitness using three measures in different time horizons in different investment portfolios. Panel A and B list machine learning mean absolute errors (MAE), calculated as $MAE_{oos} = \sum_{oos} ||\hat{\beta}_t x_{i,t+1} - y_{i,t+1}||/N$; Panel C and D list mean squared errors (MSE), calculated as $MSE_{oos} = \sum_{oos} (\hat{\beta}_t x_{i,t+1} - y_{i,t})^2/N$; and Panel E and F list variance explained (VE), calculated as $VE^2_{oos} = 1 - \sum_{oos} (\hat{\beta}_t x_{i,t+1} - y_{i,t+1})^2 / \sum_{oos} y^2_{i,t+1}$. The first row in each panel summarizes the overall out-of-sample forecasting performance; the second row provides summarization for the top 10% of all out-of-sample forecasting data; and the last row provides the forecast error for bottom 10%. The number unit in Panel A and B is in percentage (%), in C and D is in percentage squared (%^{2}); in E and F is unit-free.

Table 1: Model Fitness between Different Measures

Panel A: Comparison of Monthly MAE between Different Models

	Ridge	ENet	PLS	SGD	GB	LSVR	MLP1	MLP2	MLP3
MAE	10.44	10.32	11.28	10.28	11.08	10.18	10.59	10.56	11.13
MAE_HIGH	10.46	10.24	11.80	9.88	11.56	10.20	10.81	10.72	11.49
MAE_LOW	11.50	10.85	12.20	11.35	10.82	11.10	11.40	10.99	12.08

Panel B: Comparison of Annually MAE between Different Models

	Ridge	ENet	PLS	SGD	GB	LSVR	MLP1	MLP2	MLP3
MAE	10.36	10.22	11.19	10.18	11.00	10.08	10.50	10.48	11.05
MAE_HIGH	10.34	10.25	12.40	9.81	12.06	9.82	10.90	10.56	11.89
MAE_LOW	11.53	10.08	12.80	11.19	11.23	10.54	11.02	10.98	11.68

Panel C: Comparison of Monthly MSE between Different Models

	Ridge	ENet	PLS	SGD	GB	LSVR	MLP1	MLP2	MLP3
MSE	194.19	190.08	223.15	188.78	212.27	184.58	199.05	196.61	221.36
MSE_HIGH	195.18	189.74	234.43	175.26	226.81	184.52	200.77	200.66	229.25
MSE_LOW	232.74	207.09	260.68	224.04	205.07	217.32	234.11	216.66	265.47

Panel D: Comparison of Annually MSE between Different Models

	Ridge	ENet	PLS	SGD	GB	LSVR	MLP1	MLP2	MLP3
MSE	190.10	185.60	218.40	184.17	208.37	180.08	194.96	192.85	217.33
MSE_HIGH	193.29	188.63	257.13	171.40	239.06	163.76	198.06	190.93	242.62
MSE_LOW	232.98	179.48	285.43	218.53	225.00	199.38	221.53	219.00	255.45

Panel E: Comparison of Monthly VE between Different Models

	Ridge	ENet	PLS	SGD	GB	LSVR	MLP1	MLP2	MLP3
R2	-2.85	0.29	-26.07	0.77	-19.77	0.63	-10.29	-7.86	-19.60
R2_HIGH	-6.25	0.10	-40.02	0.54	-19.99	-2.52	-24.23	-18.76	-25.79
R2_LOW	-4.53	0.38	-38.44	0.81	-20.91	-2.87	-19.77	-16.50	-39.76

Panel F: Comparison of Annually VE between Different Models

	Ridge	ENet	PLS	SGD	GB	LSVR	MLP1	MLP2	MLP3
R2	-2.32	0.39	-18.12	0.99	-15.49	2.56	-4.95	-4.16	-16.13
R2_HIGH	-3.20	0.39	-31.99	0.83	-27.04	2.33	-14.99	-4.49	-17.49
R2_LOW	-6.35	0.70	-36.24	1.26	-12.42	-0.50	-12.04	-15.35	-28.99

4. Variable of Importance of Pricing Variables

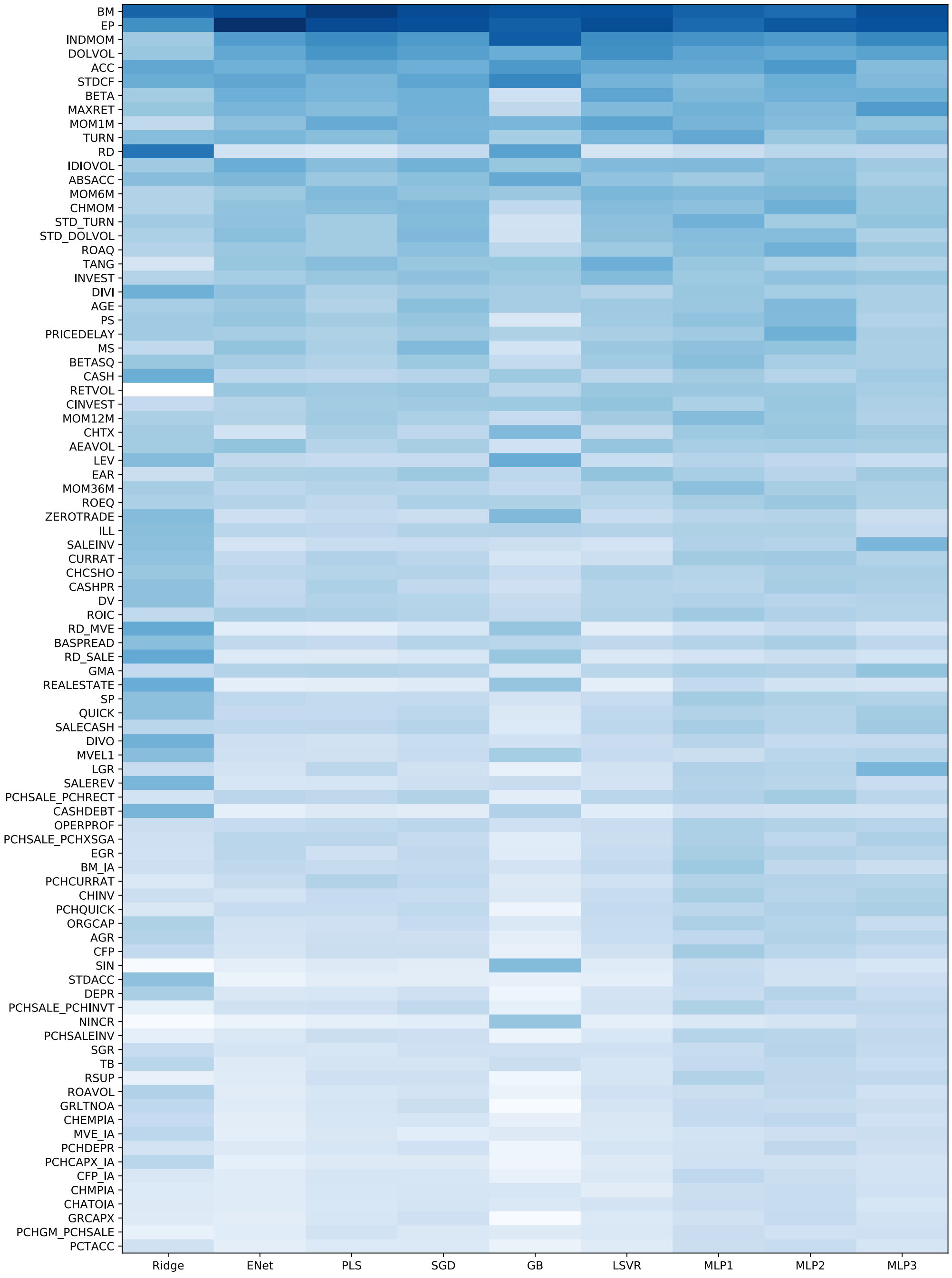


Figure 5: Variable of Importance of Pricing Variables

5. Variables of Importance of Pricing Factors by Model

Note: Figure 6 presents variables of importance of top fifteen pricing factors for each machine learning model. Variable of importance is calculated as additional variance explained by a factor compared to variance explained excluding this pricing factor in out-of-sample forecasting data. Variables of importance are adjusted to sum to one for each model.

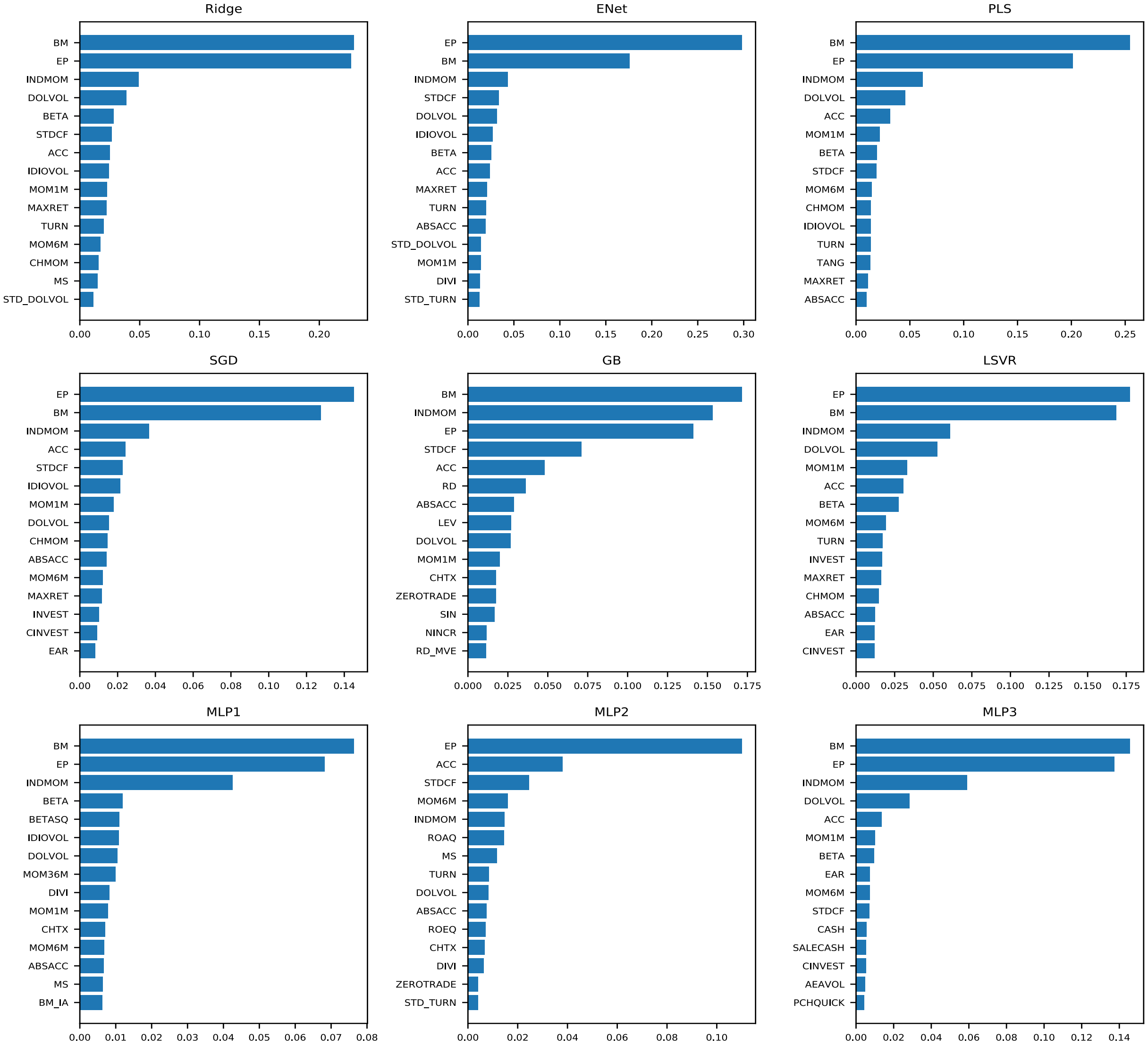


Figure 6: Variables of Importance of Characteristics

6. Variable of Importance of Macro Factors

Note: Table 2 and Figure 7 present the variable of importance of macroeconomic variables. Eight macroeconomic variables are used in each machine learning models: (1) DP, the dividend-to-price ratio; (2) EP, the earning-to-price ratio; (3) BM, the book-to-market ratio; (4) NTIS, the net equity expansion; (5) TBL, the treasure bill rate; (6) TMS, the term spread; (7) DFY, the default spread; and (8) SVAR, the stock variance. Variable of importance is calculated as additional variance explained by a factor compared to variance explained excluding this pricing factor in out-of-sample forecasting data. Variables of importance are adjusted to sum to one for each macroeconomic factor.

Table 2: Variables of Importance of Macro Factors

	Ridge	ENet	PLS	SGD	GB	LSVR	MLP1	MLP2	MLP3
BM	23.80	22.69	21.09	22.80	10.57	18.71	17.55	18.25	26.42
DFY	12.70	6.21	15.85	9.92	13.17	13.49	12.71	12.84	7.85
DP	11.27	11.65	11.13	12.25	14.32	12.93	11.28	12.69	11.96
EP	21.00	25.18	19.58	21.99	8.62	22.78	20.68	16.03	23.35
NITS	8.70	6.99	10.02	9.23	14.82	9.78	11.80	11.06	7.65
SVAR	3.47	4.85	5.67	3.83	23.21	4.88	5.45	5.53	9.38
TBL	7.96	10.38	4.11	8.30	7.31	6.51	8.39	5.61	3.66
TMS	11.11	12.05	12.56	11.70	7.98	10.92	12.14	17.98	9.72

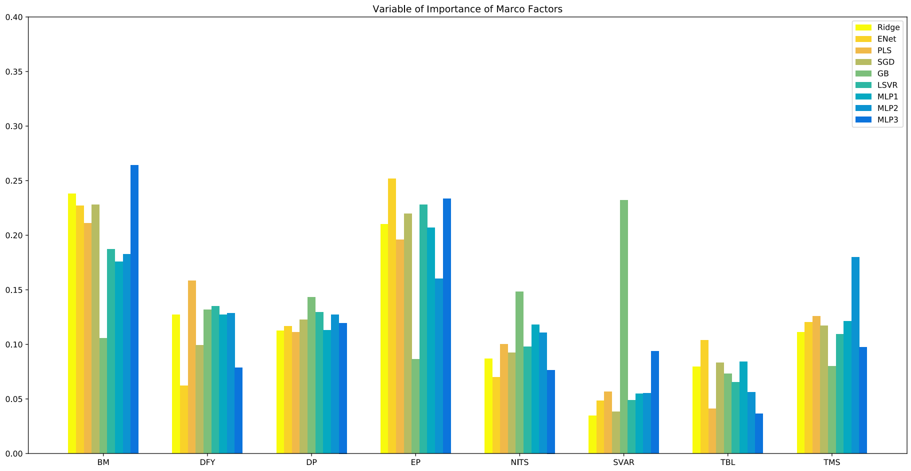


Figure 7: Variable of Importance of Macro Variables

7. Variable of Importance of Sentiment Factors

Note: Table 3 and Figure 8 present the variable of importance of sentiment variables. Nine macroeconomic variables are used in each machine learning models: (1) Sentiment of analyst report at firm-level ; (2) Sentiment of analyst report at industry level; (3) Sentiment of analyst report at market-level ; (4) Sentiment of traditional news at firm-level; (5) Sentiment of traditional news at industry level; (6) Sentiment of traditional news at market-level; (7) Sentiment of social media at firm-level; (8) Sentiment of social media at industry level; and (9) Sentiment of social media at market-level. Variable of importance is calculated as additional variance explained by a factor compared to the variance explained excluding this pricing factor in out-of-sample forecasting data. Variable of importance is adjusted to sum to one for each macroeconomic factor.

Table 3: Variables of Importance of Sentiment Factors

	Ridge	ENet	PLS	SGD	GB	LSVR	MLP1	MLP2	MLP3
AR STOCK	3.10	1.10	2.33	3.54	0.01	1.26	3.15	5.91	0.45
AR INDUSTRY	4.56	2.99	3.16	5.78	0.22	1.53	2.10	7.21	2.17
AR MARKET	13.82	9.42	19.91	10.93	21.25	7.66	11.31	8.88	16.22
TM STOCK	3.33	3.83	4.79	6.30	0.24	2.40	12.03	12.37	2.46
TM INDUSTRY	10.79	12.50	10.55	13.16	0.31	7.77	11.49	6.14	8.84
TM MARKET	41.24	33.20	45.73	32.43	63.04	55.39	32.36	35.93	44.70
SM STOCK	0.27	0.26	0.30	0.42	0.00	0.18	3.40	2.95	0.57
SM INDUSTRY	5.35	6.02	1.60	7.02	0.07	5.24	10.05	7.56	5.44
SM MARKET	17.54	30.69	11.61	20.41	14.86	18.56	14.12	13.05	19.15

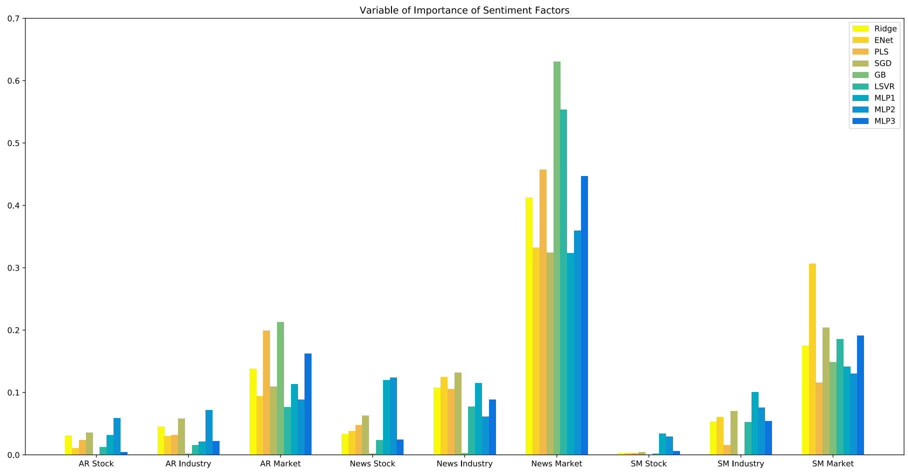


Figure 8: Variable of Importance of Sentiment Variables

8. Performance of Machine Learning Portfolios

Note: This Table report of performance of equal-weighted decile various machine learning portfolios sorted on out of sample forecasts. From top to down and left to right, I use ridge model (Ridge), ElasticNet model (ENet), Partial Least Square (PLS), Stochastic Gradient Descent (SGD), Gradient Boost(GB), Linear Support Vector Regression (LVSR), 1-layer neural network (NN1), 2-layer neural network (NN2), and 3-layers neural network (NN3) to complete the task of asset pricing. "Pred" is average predicted monthly returns, "Avg" is average realized monthly returns, "Std" is the average realized monthly return standard deviations, and "SR" is annualized Sharpe ratios.

Table 4: Performance of Machine Learning Portfolios

Ridge	Pred	Avg	Std	SR	**ENet**	Pred	Avg	Std	SR	**PLS**	Pred	Avg	Std	SR
Low(L)	-1.62	-1.20	14.22	-0.30	Low(L)	-0.04	-0.04	13.87	-0.01	Low(L)	-5.69	-1.52	14.30	-0.37
2	-0.81	-0.39	13.52	-0.10	2	-0.01	0.20	13.51	0.05	2	-3.36	-0.46	13.62	-0.12
3	-0.44	0.04	13.42	0.01	3	-0.00	0.43	13.54	0.11	3	-2.28	-0.07	13.40	-0.02
4	-0.16	0.39	13.32	0.10	4	0.01	0.41	13.36	0.10	4	-1.48	0.30	13.30	0.07
5	0.07	0.51	13.11	0.13	5	0.01	0.40	13.29	0.10	5	-0.79	0.40	13.13	0.10
6	0.29	0.70	13.07	0.18	6	0.02	0.50	13.21	0.13	6	-0.14	0.66	13.07	0.17
7	0.51	0.83	13.05	0.22	7	0.03	0.56	13.26	0.14	7	0.51	0.86	12.96	0.22
8	0.75	1.04	13.06	0.27	8	0.03	0.68	13.18	0.17	8	1.23	1.04	12.98	0.27
9	1.05	1.21	13.04	0.32	9	0.04	0.68	13.08	0.18	9	2.14	1.48	12.99	0.39
High(H)	1.66	1.44	13.36	0.37	High(H)	0.06	0.75	13.07	0.19	High(H)	4.05	1.88	13.27	0.49
H-L	3.28	2.64	13.79	0.66	H-L	0.10	0.78	13.48	0.20	H-L	9.75	3.40	13.79	0.85

SGD	Pred	Avg	Std	SR	**GB**	Pred	Avg	Std	SR	**LSVR**	Pred	Avg	Std	SR
Low(L)	-0.21	-0.72	14.39	-0.18	Low(L)	-0.29	0.02	13.11	0.00	Low(L)	-1.37	-0.70	14.35	-0.17
2	-0.12	-0.13	13.83	-0.04	2	0.06	0.37	13.11	0.09	2	-0.63	-0.12	13.67	-0.03
3	-0.07	0.17	13.52	0.04	3	0.20	0.42	13.24	0.11	3	-0.28	0.10	13.44	0.02
4	-0.03	0.35	13.40	0.09	4	0.34	0.48	13.33	0.12	4	-0.02	0.36	13.42	0.09
5	-0.00	0.53	13.28	0.13	5	0.50	0.44	13.35	0.11	5	0.21	0.46	13.30	0.12
6	0.03	0.71	13.16	0.18	6	0.65	0.41	13.31	0.10	6	0.42	0.62	13.06	0.16
7	0.05	0.75	12.95	0.19	7	0.78	0.52	13.45	0.13	7	0.64	0.74	13.13	0.19
8	0.09	0.89	12.91	0.23	8	0.94	0.62	13.51	0.16	8	0.90	0.82	13.01	0.21
9	0.13	0.98	12.87	0.26	9	1.13	0.61	13.49	0.15	9	1.21	1.06	12.93	0.28
High(H)	0.20	1.03	12.92	0.27	High(H)	1.62	0.68	13.48	0.17	High(H)	1.91	1.22	12.94	0.32
H-L	0.42	1.75	13.67	0.44	H-L	1.90	0.66	13.30	0.17	H-L	3.28	1.92	13.66	0.49

NN1	Pred	Avg	Std	SR	**NN2**	Pred	Avg	Std	SR	**NN3**	Pred	Avg	Std	SR
Low(L)	0.01	0.12	13.75	0.03	Low(L)	0.17	0.01	13.41	-0.00	Low(L)	-1.00	-0.08	13.57	-0.03
2	0.54	0.38	13.50	0.09	2	0.65	0.33	13.27	0.08	2	-0.11	0.25	13.33	0.06
3	0.82	0.50	13.33	0.13	3	0.89	0.48	13.41	0.12	3	0.34	0.40	13.26	0.10
4	1.03	0.50	13.16	0.13	4	1.06	0.49	13.36	0.12	4	0.69	0.49	13.23	0.12
5	1.20	0.57	13.22	0.14	5	1.21	0.44	13.33	0.11	5	1.00	0.51	13.22	0.13
6	1.36	0.48	13.12	0.12	6	1.34	0.56	13.38	0.14	6	1.30	0.52	13.34	0.13
7	1.52	0.59	13.25	0.15	7	1.50	0.59	13.28	0.15	7	1.60	0.63	13.28	0.16
8	1.70	0.53	13.18	0.14	8	1.70	0.54	13.25	0.14	8	1.92	0.65	13.32	0.16
9	1.92	0.46	13.32	0.12	9	1.93	0.57	13.29	0.14	9	2.32	0.65	13.38	0.16
High(H)	2.40	0.44	13.57	0.11	High(H)	2.29	0.55	13.40	0.14	High(H)	3.07	0.56	13.46	0.14
H-L	2.39	0.32	13.66	0.08	H-L	2.12	0.54	13.41	0.14	H-L	4.07	0.64	13.51	0.17

9. Cumulative Long-Short Returns by Different Models

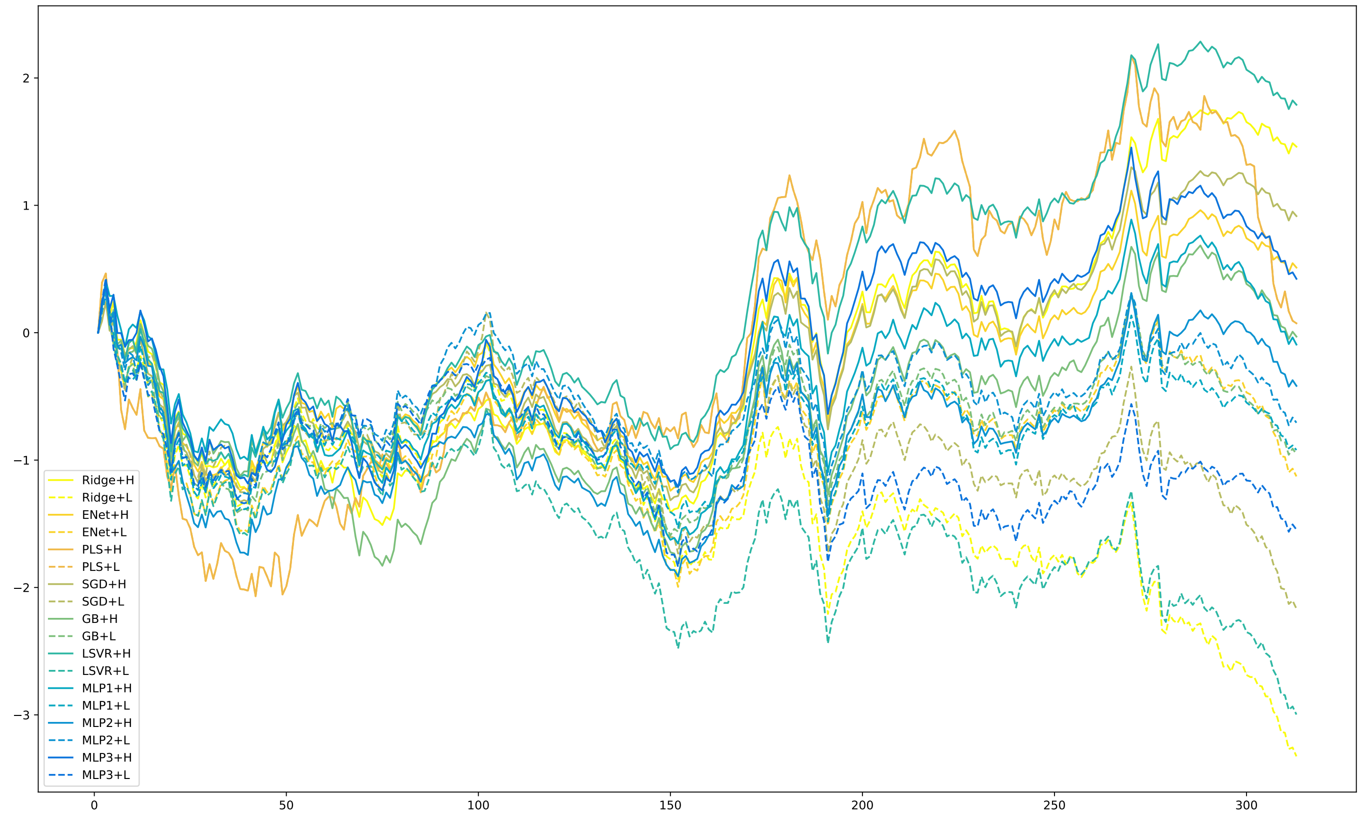


Figure 9: Log-Returns of Different Classifiers

10. List of Pricing Factors in Literature

Table 5: Firm Characteristics and Corresponding CSMAR Files

No.	Variables	Authors	Year	Paper	Freqency	CSMAR
1	ACC	Bandyopadhyay, Huang & Wirjanto	2010	WP	Annually	FS
2	ABSACC	Bandyopadhyay, Huang & Wirjanto	2010	WP	Annually	FS
3	AEAVOL	Lerman, Livnat & Mendenhall	2007	WP	Quarterly	TRD_Dalyr, IAR_Rept
4	AGE	Jiang, Lee & Zhang	2005	RAS	Annually	FS
5	AGR	Cooper, Gulen & Schill	2008	JF	Annually	FS
6	BASPREAD	Amihud & Mendelson	1989	JF	Monthly	STK_MKT_Dalyr
7	BETA	Fama & MacBeth	1973	JPE	Monthly	STK_MKT_Stkbtal
8	BETASQ	Fama & MacBeth	1973	JPE	Monthly	STK_MKT_Stkbtal
9	BM	Rosenberg, Reid & Lanstein	1985	JPM	Annually	TRD_Mnth, FS
10	BM_IA	Rosenberg, Reid & Lanstein	1985	JPM	Annually	TRD_Mnth, TRD_Co, FS
11	CASH	Palazzo	2012	JFE	Quarterly	FS
12	CASHDEBT	Ow and Penman	1989	JAE	Annually	FS
13	CAHSPR	Chandrashekar & Rao	2009	WP	Annually	TRD_Mnth, FS
14	CFP	Desai, Rajgopal & Venkatachalam	2004	TAR	Annually	TRD_Mnth, FS
15	CFP_IA	Desai, Rajgopal & Venkatachalam	2004	TAR	Annually	TRD_Mnth, FS, TRD_Co
16	CHATOIA	Soliman	2008	TAR	Annually	FS, TRD_Co
17	CHCSHO	Pontiff & Woodgate	2008	JF	Annually	TRD_Mnth
18	CHEMP	Asness, Porter & Stevens	1994	WP	Annually	TRD_Co, CG_Ybasic
19	CHINV	Thomas & Zhang	2002	RAS	Annually	FS
20	CHMOM	Gettleman & Marks	2006	WP	Monthly	TRD_Mnth
21	CHPMIA	Soliman	2008	TAR	Annually	TRD_Co, FS
22	CHTX	Thomas & Zhang	2001	JAR	Quarterly	FS
23	CINVEST	Titman, Wei & Xie	2004	JFQA	Quarterly	FS
24	CONVIND	Valta	2016	JFQA	Annually	-
25	CURRAT	Ou & Penman	1989	JAE	Annually	FS
26	DEPR	Holthausen & Larcker	1992	JAE	Annually	FS
27	DIVI	Michaely, Thaler & Womack Michaely	1995	JF	Annually	CD_Dividend
28	DIVO	Michaely, Thaler & Womack Michaely	1995	JF	Annually	CD_Dividend
29	DOLVOL	Chordia, Subrahmanyam & Anshuman	2001	JFE	Annually	TRD_Mnth
30	PV	Litzenberger & Ramaswamy	1982	JF	Annually	CD_Dividend
31	EAR	Kishore, Brandt, Santa-Clara & Venkatachalam	2008	WP	Quarterly	TRD_Mnth, FS

Table 5: Firm Characteristics and Corresponding CSMAR Files

No.	Variables	Authors	Year	Paper	Freqency	CSMAR
32	EGR	Richardson, Sloan, Soliman & Tuna	2005	JAE	Annually	FS
33	EP	Basu	1977	JF	Annually	TRD_Mnth, FS
34	GMA	Novy-Marx	2013	JFE	Annually	FS
35	GRCAPX	Anderson & Garcia-Feijoo 2006, JF	2006	JF	Annually	FS
36	GRLTNOA	Fairfield, Whisenant & Yohn	2003	TAR	Annually	FS
37	HERF	Hou & Robinson	2006	JF	Annually	TRD_Co, FS
38	HIRE	Bazdresch, Belo & Lin	2014	JPE	Annually	CG_Ybasic
39	IDIOVOL	Ali, Hwang & Trombley	2003	JFE	Monthly	STK_MKT_Stkbtal
40	ILL	Amihud	2002	JFM	Monthly	TRD_Dalyr
41	INDMOM	Moskowitz & Grinblatt	1999	JF	Monthly	TRD_Mnth, TRD_Co, FS
42	INVEST	Chen & Zhang	2010	JF	Annually	FS
43	LEV	Bhandari	1988	JF	Annually	FS, TRD_Mnth
44	LGR	Richardson, Sloan, Soliman & Tuna	2005	JAE	Annually	FS
45	MAXRET	Bali, Cakici & Whitelaw	2011	JFE	Monthly	TRD_Dalyr
46	MOM12M	Jegadeesh	1990	JF	Monthly	TRD_Mnth
47	MOM1M	Jegadeesh & Titman	1993	JF	Monthly	TRD_Mnth
48	MOM36M	Jegadeesh & Titman	1993	JF	Monthly	TRD_Mnth
49	MOM6M	Jegadeesh & Titman	1993	JF	Monthly	TRD_Mnth
50	MS	Mohanram	2005	RAS	Quarterly	FS
51	MVEL1	Banz	1981	JFE	Monthly	TRD_Mnth
52	MVE_IA	Asness, Porter & Stevens	2000	WP	Annually	TRD_Co, TRD_Mnth
53	NINCR	Barth, Elliott and Finn	1999	JAR	Quarterly	FS
54	OPERPROF	Fama and French	2015	JFE	Annually	FS
55	ORGCAP	Eisfeldt & Papanikolaou	2013	JF	Annually	FS, CPI
56	PCHCAPX_IA	Abarbanell & Bushee	1998	TAR	Annually	FS, TRD_Co
57	PCHCURRAT	Ou & Penman	1989	JAE	Annually	FS
58	PCHDEPR	Holthausen & Larcker	1992	JAE	Annually	FS
59	PCHGM_PCHSALE	Abarbanell & Bushee	1998	TAR	Annually	FS
60	PCHQUICK	Ou & Penman	1989	JAE	Annually	FS
61	PCHSALE_PCHINVT	Abarbanell & Bushee	1998	TAR	Annually	FS
62	PCHSALE_PCHRECT	Abarbanell & Bushee	1998	TAR	Annually	FS
63	PCHSALE_PCHXSGA	Abarbanell & Bushee	1998	TAR	Annually	FS
64	PCHSALEINV	Ou & Penman	1989	JAE	Annually	FS

Table 5: Firm Characteristics and Corresponding CSMAR Files

No.	Variables	Authors	Year	Paper	Freqency	CSMAR
65	PCTACC	Hafzalla, Lundholm & Van Winkle	2011	TAR	Annually	FS
66	PRICEDELAY	Hou & Moskowitz	2005	RFS	Monthly	TRD_Weekcm
67	PS	Piotroski	2000	JAR	Annually	FS
68	QUICK	Ou & Penman	1989	JAE	Annually	FS
69	RD	Eberhart, Maxwell & Siddique	2004	JF	Annually	FS
70	RD_MVE	Guo, Lev & Shi	2006	JBFA	Annually	FS, TRD_Mnth
71	RD_SALE	Guo, Lev & Shi 2006, JBFA	2006	JBFA	Annually	FS
72	REALESTATE	Tuzel	2010	RFS	Annually	FS
73	RETVOL	Ang, Hodrick, Xing & Zhang	2006	JF	Monthly	STK_MKT_ThrfacDay, TRD_Dalyr
74	ROAQ	Balakrishnan, Bartov & Faurel	2010	JAE	Quarterly	FS
75	ROAVOL	Francis, LaFond, Olsson & Schipper	2004	TAR	Quarterly	FS
76	ROEQ	Hou, Xue & Zhang	2015	RFS	Quarterly	FS
77	ROIC	Brown & Rowe	2007	WP	Annually	FS
78	RSUP	Kama	2009	JBFA	Quarterly	FS
79	SALECASH	Ou & Penman	1989	JAE	Annually	FS
80	SALEINV	Ou & Penman	1989	JAE	Annually	FS
81	SALEREV	Ou & Penman	1989	JAE	Annually	FS
82	SECURED	Valta	2016	JFQA	Annually	-
83	SECUREDIND	Valta	2016	JFQA	Annually	-
84	SGR	Lakonishok, Shleifer & Vishny	1994	JF	Annually	FS
85	SIN	Hong & Kacperczyk	2009	JFE	Annually	TRD_Co
86	SP	Barbee, Mukherji, & Raines	1996	FAJ	Annually	TRD_Mnth, FS
87	STD_DOLVOL	Chordia, Subrahmanyam & Anshuman	2001	JFE	Monthly	TRD_Mnth
88	STD_TURN	Chordia, Subrahmanyam & Anshuman	2001	JFE	Monthly	TRD_Mnth
89	STDACC	Bandyopadhyay, Huang & Wirjanto	2010	WP	Quarterly	FS
90	STDCF	Huang	2009	JEF	Quarterly	FS
91	TANG	Almeida & Campello	2007	RFS	Quarterly	FS
92	TB	Lev & Nissim	2004	TAR	Annually	FS
93	TURN	Datar, Naik & Radcliffe	1998	JFM	Annually	TRD_Mnth
94	ZEROTRADE	Liu	2006	JFE	Monthly	TRD_Mnth

11. Detail of Calculation Method of Pricing Factors

Table 6: Details of Characteristics and Items in CSMAR Files

No.	Variables	Description	Items in CSMAR
1	ACC	Total accruals, where total accruals are changes in working capital minus depreciation and amortization	a001100000, a001101000, a002100000, a002125000, a002113000, b001212000
2	ABSACC	Absolute value of total accruals, where total accruals are changes in working capital minus depreciation and amortization	a001100000, a001101000, a002100000, a002125000, a002113000, b001212000
3	AEAVOL	Abnormal earnings announcement volume, which is calculated as average of (-1, 1) trading volume divided by average of (-63, -8) trading volume minus one	dnshrtrd
4	AGE	Year since first CSMAR coverage	accper
5	AGR	Asset growth (AGR) is the 1-year percentage change in total firm assets	a001000000
6	BASPREAD	Bid-Ask Spread are not recorded, we use directly use 'liquidity' measure from CSMAR risk factor database instead	liquidity
7	BETA	Beta	beta2
8	BETASQ	Beta Squared	beta2
9	BM	Book-to-market Ratio	msmvosd, a003000000
10	BM_IA	Industry-adjusted Book-to-market Ratio	msmvosd, nnindnme, a003000000
11	CASH	Cash Holdings	a001101000
12	CASHDEBT	Cash flow to total debt	c001000000, a002000000
13	CAHSPR	Cash productivity, calculated as market value of equity plus book value of debt minus total physical assets divided by cash, (MV - TPA)/Cash	msmvosd, a002000000, a001000000, a001101000
14	CFP	Cash flow to price ratio, which is cash flow from operation scaled by market value of equity	c001000000, msmvosd,
15	CFP_IA	Industry-adjusted cash flow to price ratio, which is cash flow from operation scaled by market value of equity minus industry average	c001000000, nnindnme, msmvosd
16	CHATOIA	Industry-adjusted change in asset turnover; Assets turnover, ATO, is calculated as Sales divided by Net Operating Assets (NOA)	nnindnme, a003200000, a001000000, a001101000, a001000000, a002000000, a003101000, a003200000, b001100000
17	CHCSHO	Changes in Shares Outstanding, calculated as the natural logarithm of Shares Outstanding at time t minus natural logarithm of shares outstanding at time t-1	msmvosd, mclsprc
18	CHEMP	Industry adjusted change in employees	nnindnme, y0601b
19	CHINV	Change in total inventory deflated by average total assets	a001123000, a001000000
20	CHMOM	change in 6-month momentum, which is an acceleration strategy	mretwd
21	CHPMIA	Industry Adjusted Change in Profit Margin, which is calculated as operating income divided by total sales	nnindnme, b001300000, b001100000
22	CHTX	Change in tax expense, measured as tax expense per share in quarter Q minus EPS in quarter Q-4, scaled by asset per share in Quarter t-4	b002100000, a001000000

Table 6: Details of Characteristics and Items in CSMAR Files

No.	Variables	Description	Items in CSMAR
23	CINVEST	Corporate Investment, measured by corporate expenditure at time t divided by the average of that in t-1, t-2 and t-3	c002003000, c001022000, c002003000, a001000000
24	CONVIND	This measure is not derivable from data available in China financial market	-
25	CURRAT	Current Ratio, measured by Current Asset divided by Current Liability	a001100000, a002100000
26	DEPR	Depreciation ratio, measured by Depreciation divided by PP&E	d000103000, a001212000
27	DIVI	Dividend Initiation	btperdiv
28	DIVO	Dividend Omission	btperdiv
29	DOLVOL	Dollar Trading Volume, measured by the natural logarithm of the trading volume	mnvaltrd
30	PV	Dividend to price	btperdiv, price1
31	EAR	Earning Announcement Return, measured by the abnormal FF return for firm i in quarter q recorded over a three-day window centered on the announcement date	msmvosd, a003000000, a001000000, dretwd, dsmvosd
32	EGR	Growth in common shareholder equity	a003101000, a001000000
33	EP	Earnings to Price	mclsprc, msmvosd, b003000000, b002000000
34	GMA	Gross Profit Margin	b001100000, b001200000, a001000000
35	GRCAPX	Growth in capital expenditure	c002003000, c001022000, c002003000, a001000000
36	GRLTNOA	Growth in long term net operating assets, which is measured by growth of NOA minus of growth of working capital	a001111000, a001123000, a001125000, a001212000, a001218000, a001223000, a002108000, a002126000, a002209000, a001000000, d000103000
37	HERF	Industry Sales Concentration, which is measured by Herfindahl Index for each industry	nnindnme, b001100000
38	HIRE	Employee growth rate/Labor hire growth rate	y0601b
39	IDIOVOL	Idiosyncratic return volatility, which is measured by volatiles index in CSMAR, which calucalte the variance of return residual by consecutively regression 250 days	volatility
40	ILL	Illiquidity, measured by the ratio of the sum of the daily volume to the sum of the absolute return	dretwd, dnvaltrd
41	INDMOM	Industry Momentum	trdmnt, msmvosd, nnindnme, a003000000, a001000000
42	INVEST	Capital expenditures and inventory, which is defined by annual change in gross property, plant, and equipment plus the annual change in inventories divided by the lagged book value of assets	a001212000, a001123000, a001000000
43	LEV	Debt to equity ratio, which is book value of total assets minus book value of common equity scaled by market value of common equity	a001000000, a003101000, msmvttl
44	LGR	Growth in long-term debt	a002206000, a001000000, a001000000
45	MAXRET	Maximum daily return	dretwd
46	MOM12M	12 month momentum	mretwd
47	MOM1M	1 month momentum	mretwd
48	MOM36M	36 month momentum	mretwd
49	MOM6M	6 month momentum	mretwd

Table 6: Details of Characteristics and Items in CSMAR Files

No.	Variables	Description	Items in CSMAR
50	MS	Financial statement score, sum of 8 binary financial indicators	b002000000, a001000000, d000100000, b001100000, a001219000, c002003000, c001022000, c002003000, b001209000
51	MVEL1	market value of equity, which is measured by market value of equity minus average value of equity scaled by the average value of equity	msmvttl
52	MVE_IA	industry adjusted size	msmvttl, nnindnme
53	NINCR	Number of earnings increase in past five years, which is a indicator variable that equals one for firm years least five consecutive prior years of increasing earnings, and zero otherwise	b002000000
54	OPERPROF	Operating Profitability, measured by annual revenues minus cost of goods sold, interest expense, and selling, general, and administrative expenses, all divided by book equity at the end of fiscal year t-1.	b001201000, b001209000, bbd1102203, b001210000, b001211000, a003000000
55	ORGCAP	Organizational Capital, which is measured by d times lagged ORGCAP and current SG&A deflated by CPI	b001209000, b001210000, b001211000, b0f1208000, a001000000, CPI
56	PCHCAPX_IA	Industry adjusted % change in capital expenditure	c002003000, c001022000, c002003000, a001000000, nnindnme
57	PCHCURRAT	% change in current ratio	a001100000, a002100000
58	PCHDEPR	% change in depreciation	d000103000
59	PCHGM_PCHSALE	% change in gross margin - % change in sales	b001100000, b001200000
60	PCHQUICK	% change in quick ratio	a001101000, a001107000, a001111000, a002100000
61	PCHSALE_PCHINVT	% change in sales - % change in inventory	b001100000, a001123000
62	PCHSALE_PCHRECT	% change in sales - % change in A/R	b001100000, a001111000
63	PCHSALE_PCHXSGA	% change in sales - % change in SG&A	b001209000, b001210000, b001211000, b001100000
64	PCHSALEINV	% change sales-to-inventory	b001100000, a001123000
65	PCTACC	Percentage of operating accruals	b002000000, d000100000
66	PRICEDELAY	Price Delay, measured by deriving regression 1-R1/R2	cwretwdos
67	PS	F-score, sum of nine binary financial report indicators	b002000000, a001000000, c001000000, a002206000, a001100000, a002100000, a003102000, b001100000, b001200000
68	QUICK	Quick Ratio	a001101000, a001107000, a001111000, a002100000
69	RD	R&D Increase	a001219000
70	RD_MVE	R&D to market cap	a001219000, msmvttl
71	RD_SALE	R&D to sales	a001219000, b001100000
72	REALESTATE	Real estate holding	a001211000, a001212000
73	RETVOL	Return Volatility	riskpremium1, smb1, hml1, dretwd
74	ROAQ	Return on Assets	b002000000, a001000000
75	ROAVOL	Earnings volatility, which is measured by the devision between standard deviation of NIBE and standard deviation of CFO	b002000000, c001000000
76	ROEQ	Return on Equity	b002000000, a003000000
77	ROIC	Return on invested capital, which is measured by operating income devided by book value of invested capital	b002000000, a004000000, a001101000
78	RSUP	Revenue Surprise	b001100000, a003101000

Table 6: Details of Characteristics and Items in CSMAR Files

No.	Variables	Description	Items in CSMAR
79	SALECASH	Sales to Cash	b001100000, a001101000
80	SALEINV	Sales to inventory	b001100000, a001123000
81	SALEREV	Sales to Receivables	b001100000, a001111000
82	SECURED	Secured debt	-
83	SECUREDIND	Secured debt indicator	-
84	SGR	Sales growth	b001100000
85	SIN	Binary indicator for Tobacco, Alcohol and Gaming firms	nnindcd
86	SP	Sales to price	msmvosd, b001100000
87	STD_DOLVOL	Volatility of liquidity (dollar trading volume)	mnvaltrd
88	STD_TURN	Volatility of liquidity (share turnover)	mnvaltrd, msmvosd
89	STDACC	Accrual volatility	a001100000, a001101000, a002100000, a002125000, a002113000, b001212000
90	STDCF	Cash flow volatility, measured as the rolling standard deviation of the standardized cash-flow over the past sixteen quarters (four years), scaled by sales	c001000000, b001100000
91	TANG	Tangibility = 0:715*Receivables + 0:547*Inventory + 0:535*Capital	a001111000, a001123000, a001212000, a001101000, a001000000
92	TB	Tax income to book income	b002100000, b002000000
93	TURN	Share turnover	mnshrtrd, msmvosd
94	ZEROTRADE	Zero trading days	ndaytrd

12. Source of Sentiment Factors and Related Literature

Table 7: Details of Sentiment Factors

No.	Variables	Description	Source
95	AR STOCK	Textual Sentiment of Analyst Reports of Stock	Tencent Finance
96	AR INDUSTRY	Textual Sentiment of Analyst Reports of Industry	Tencent Finance
97	AR MARKET	Textual Sentiment of Analyst Reports of Market	Tencent Finance
98	TM STOCK	Textual Sentiment of Traditional News of Stock	DataGo
99	TM INDUSTRY	Textual Sentiment of Traditional News of Industry	DataGo
100	TM MARKET	Textual Sentiment of Traditional News of Market	DataGo
101	SM STOCK	Textual Sentiment of Social Media of Stock	Eastmoney.com
102	SM INDUSTRY	Textual Sentiment of Social Media of Industry	Eastmoney.com
103	SM MARKET	Textual Sentiment of Social Media of Market	Eastmoney.com

Table 8: Literatures of Used Sentiment Factors

No.	Variables	Description	Source
1	OPN_{AR}	Sentiment of Analyst Reports	Huang et.al., 2014
2	OPN_{TM}	Sentiment of Traditional Media Articles	Piotroski et.al.,2016
3	OPN_{SM}	Sentiment of Social Media	Wang et.al., 2019

13. Summary Statistics of Pricing Factors

Table 9: Detailed Summary Statistics of Pricing Factors

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	N	Mean	SD	P25	P50	P75
ACC	447,453.0000	0.0057	0.3806	-0.1436	-0.0176	0.0952
ABSACC	447,453.0000	0.0043	0.3076	-0.1836	-0.1168	0.0106
AEAVOL	447,453.0000	-0.0039	0.1790	-0.1106	-0.0537	0.0368
AGE	447,453.0000	0.0393	0.4056	-0.2669	-0.0669	0.3332
AGR	447,453.0000	-0.0019	0.1529	-0.0684	-0.0328	0.0150
BASPREAD	447,453.0000	-0.0082	0.0938	-0.0272	-0.0259	-0.0208
BETA	447,453.0000	0.0669	0.3097	-0.1249	0.0763	0.2609
BETASQ	447,453.0000	0.0360	0.2621	-0.1449	0.0091	0.1763
BM	447,453.0000	-0.0065	0.1180	-0.0449	-0.0326	-0.0119
BM_IA	447,453.0000	0.0021	0.1226	-0.0301	-0.0095	0.0070
CASH	447,453.0000	0.0045	0.1385	-0.0418	-0.0333	-0.0119
CASHDEBT	447,453.0000	-0.0014	0.1247	0.0178	0.0268	0.0296
CASHPR	447,453.0000	0.0013	0.1439	-0.0544	-0.0381	-0.0028
CFP	447,453.0000	0.0001	0.1479	-0.0449	-0.0287	0.0050
CFP_IA	447,453.0000	0.0007	0.1582	-0.0444	-0.0072	0.0207
CHATOIA	447,453.0000	-0.0006	0.2058	-0.0135	0.0070	0.0229
CHCSHO	447,453.0000	0.0004	0.2435	-0.1371	-0.1370	0.0535
CHEMPIA	447,453.0000	0.0000	0.1563	-0.0007	0.0220	0.0365
CHINV	447,453.0000	-0.0019	0.2196	-0.0842	-0.0390	0.0460
CHMOM	447,453.0000	-0.0020	0.2950	-0.1445	-0.0004	0.1466
CHMPIA	447,453.0000	0.0010	0.1341	-0.0174	-0.0104	-0.0027
CHTX	447,453.0000	-0.0036	0.1902	-0.0680	-0.0363	0.0343
CINVEST	447,453.0000	0.0000	0.2274	-0.1751	-0.0200	0.1232
CURRAT	447,453.0000	0.0015	0.1793	-0.0891	-0.0512	0.0132
DEPR	447,453.0000	0.0035	0.1668	-0.0838	-0.0320	0.0304
DIVI	447,453.0000	-0.0120	0.2300	-0.0643	-0.0643	-0.0643
DIVO	447,453.0000	-0.0065	0.3023	-0.0973	-0.0973	-0.0973
DOLVOL	447,453.0000	0.0335	0.3267	-0.1622	0.0695	0.2610
DV	447,453.0000	-0.0021	0.2309	-0.1673	-0.0883	0.0730
EAR	447,453.0000	0.0035	0.2960	-0.1536	-0.0080	0.1386
EGR	447,453.0000	-0.0003	0.2317	-0.0824	-0.0251	0.0430

EP	447,453.0000	-0.0123	0.1436	-0.0460	-0.0239	0.0091
GMA	447,453.0000	-0.0047	0.2524	-0.0980	0.0028	0.1207
GRCAPX	447,453.0000	0.0011	0.2758	-0.1020	0.0028	0.0991
GRLTNOA	447,453.0000	-0.0025	0.2266	-0.1234	-0.0607	0.0581
IDIOVOL	447,453.0000	-0.0230	0.3496	-0.2739	-0.0866	0.1745
ILL	447,453.0000	-0.0098	0.0857	-0.0289	-0.0272	-0.0210
INDMOM	447,453.0000	-0.0017	0.3167	-0.1897	-0.0084	0.1468
INVEST	447,453.0000	-0.0052	0.1983	-0.1016	-0.0480	0.0417
LEV	447,453.0000	-0.0036	0.1293	-0.0449	-0.0355	-0.0158
LGR	447,453.0000	0.0023	0.2359	-0.0587	-0.0465	0.0210
MAXRET	447,453.0000	-0.0146	0.3793	-0.3197	-0.0998	0.2617
MOM1M	447,453.0000	-0.0011	0.2807	-0.1690	-0.0215	0.1396
MOM6M	447,453.0000	-0.0036	0.2337	-0.1481	-0.0408	0.0727
MOM12M	447,453.0000	-0.0015	0.2214	-0.1322	-0.0519	0.0467
MOM36M	447,453.0000	0.0002	0.1976	-0.0658	-0.0658	0.0113
MS	447,453.0000	0.0355	0.4302	-0.1355	0.1484	0.4323
MVEL1	447,453.0000	-0.0019	0.1601	-0.0739	-0.0504	-0.0019
MVE_IA	447,453.0000	-0.0009	0.0768	-0.0130	0.0140	0.0270
NINCR	447,453.0000	0.0013	0.1201	-0.0131	-0.0131	-0.0131
OPERPROF	447,453.0000	-0.0029	0.2111	-0.0750	-0.0003	0.0804
ORGCAP	447,453.0000	0.0093	0.2301	-0.1396	-0.0456	0.0797
PCHCAPX_IA	447,453.0000	0.0002	0.1827	-0.0875	-0.0381	0.0275
PCHCURRAT	447,453.0000	0.0029	0.3084	-0.1237	0.0108	0.1289
PCHDEPR	447,453.0000	-0.0014	0.2718	-0.1205	-0.0355	0.0943
PCHGM_PCHSALE	447,453.0000	-0.0004	0.2106	-0.0356	-0.0019	0.0293
PCHQUICK	447,453.0000	0.0004	0.3216	-0.1527	0.0002	0.1487
PCHSALE_PCHINVT	447,453.0000	0.0033	0.2780	-0.1153	0.0021	0.1181
PCHSALE_PCHRECT	447,453.0000	0.0058	0.2923	-0.1220	-0.0083	0.1181
PCHSALE_PCHXSGA	447,453.0000	0.0063	0.2034	-0.0567	0.0127	0.0747
PCHSALEINV	447,453.0000	0.0032	0.2941	-0.1214	0.0046	0.1275
PCTACC	447,453.0000	-0.0013	0.1831	-0.0202	0.0197	0.0485
PRICEDELAY	447,453.0000	0.0023	0.2917	-0.2106	-0.0956	0.1155
PS	447,453.0000	-0.0127	0.3869	-0.2733	-0.0311	0.2111
QUICK	447,453.0000	0.0007	0.1745	-0.0868	-0.0520	0.0066
RD	447,453.0000	0.0022	0.1771	-0.0217	-0.0217	-0.0217
RD_MVE	447,453.0000	0.0028	0.1389	-0.0282	-0.0282	-0.0282
RD_SALE	447,453.0000	0.0031	0.1451	-0.0305	-0.0305	-0.0305
REALESTATE	447,453.0000	0.0023	0.1269	-0.0231	-0.0231	-0.0223

RETVOL	447,453.0000	-0.0081	0.2861	-0.2213	-0.0684	0.1470
ROAQ	447,453.0000	-0.0155	0.2742	-0.1537	-0.0574	0.1019
ROAVOL	447,453.0000	-0.0034	0.1616	-0.0899	-0.0422	0.0215
ROEQ	447,453.0000	-0.0100	0.2120	-0.0949	-0.0263	0.0745
ROIC	447,453.0000	-0.0044	0.2602	-0.1096	-0.0201	0.0965
RSUP	447,453.0000	-0.0000	0.2542	-0.0744	0.0000	0.0974
SALECASH	447,453.0000	-0.0130	0.2356	-0.1724	-0.0903	0.0524
SALEINV	447,453.0000	0.0052	0.2332	-0.1169	-0.0736	0.0020
SALEREV	447,453.0000	0.0069	0.1855	-0.0714	-0.0580	-0.0168
SGR	447,453.0000	0.0003	0.1616	-0.0689	-0.0229	0.0274
SIN	447,453.0000	-0.0002	0.1350	-0.0184	-0.0184	-0.0184
SP	447,453.0000	-0.0047	0.1404	-0.0591	-0.0452	-0.0142
STD_DOLVOL	447,453.0000	0.0319	0.3374	-0.1827	0.0581	0.2736
STD_TURN	447,453.0000	-0.0241	0.3682	-0.2571	-0.0083	0.2317
STDACC	447,453.0000	0.0002	0.1174	-0.0299	-0.0243	-0.0110
STDCF	447,453.0000	0.0130	0.5975	-0.4734	-0.1420	0.8764
TANG	447,453.0000	0.0197	0.3323	-0.1649	0.0401	0.2088
TB	447,453.0000	0.0039	0.1832	-0.0751	-0.0328	0.0476
TURN	447,453.0000	-0.0043	0.2113	-0.1395	-0.0781	0.0455
ZEROTRADE	447,453.0000	-0.0019	0.1462	-0.0402	-0.0402	-0.0402
AR STOCK	447,453.0000	0.0000	0.1843	-0.0445	0.0000	0.1055
AR INDUSTRY	447,453.0000	0.0000	0.2756	-0.1164	0.0000	0.1979
AR MARKET	447,453.0000	0.0000	0.2812	-0.0758	0.0000	0.1992
TM STOCK	447,453.0000	-0.0016	0.2850	-0.1168	0.0183	0.1715
TM INDUSTRY	447,453.0000	-0.0004	0.2853	-0.1441	0.0000	0.1947
TM MARKET	447,453.0000	-0.0014	0.2851	-0.2029	0.0000	0.2323
SM STOCK	447,453.0000	-0.0000	0.1649	-0.0460	0.0000	0.0114
SM INDUSTRY	447,453.0000	-0.0000	0.2585	-0.1050	0.0000	0.0519
SM MARKET	447,453.0000	-0.0000	0.3068	-0.1142	0.0000	0.0613
DP	447,453.0000	0.0148	0.3460	-0.2634	-0.0334	0.3176
EP_macro	447,453.0000	0.0281	0.4583	-0.2842	-0.0191	0.3565
BM_macro	447,453.0000	0.0780	0.4249	-0.1852	0.1531	0.4295
NITS	447,453.0000	-0.0275	0.1919	-0.1270	-0.0589	0.0209
TBL	447,453.0000	-0.0519	0.0888	-0.1302	-0.0689	0.0229
TMS	447,453.0000	0.0000	0.3524	-0.2775	0.0000	0.1690
DFY	447,453.0000	-0.0211	0.3856	-0.2237	0.0000	0.0831
SVAR	447,453.0000	-0.0159	0.0523	-0.0393	-0.0322	-0.0160