Expert Systems With Applications 196 (2022) 116629

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

How to make machine select stocks like fund managers? Use scoring and screening model☆

Yanrui Li [a], Kaiyou Fu [b], Yuchen Zhao [a], Chunjie Yang [a,*]

[a] College of Control Science and Engineering, State Key Laboratory of Industrial Control Technology, Zhejiang University, Zheda Road 38, Hangzhou, China
[b] School of Medicine, Zhejiang University, Hangzhou, China

ARTICLE INFO

Keywords:
Stock selection
Model fusion
Stock screening
Factor modeling
Deep learning

ABSTRACT

With the development of technology and the abundance of data, many novel methods like artificial intelligence and machine learning have emerged for quantitative finance. This work tries to build a framework with screening function to help investors create a portfolio of stocks based on data from multiple sources, including historical trading data, factor data, financial data and media data. The framework integrates scoring and screening models. The scoring model consists of Seq2Seq model using historical trading data and a factor model using a new bottom-up discretization method while the screening model is composed of a novel discriminative model and a media model based on the weighted stock relation graph. Two types of model are fused to select portfolio with screening ability. This framework has been verified in China's A-share market, and is proved to be effective. We also noticed that the fused model is sensitive to the scale of selected stocks and the length of prediction period, which means it can be quickly adjusted according to our trading strategy.

1. Introduction

Quantitative trading has been used in various fields of finance, including financial instruments selection, timing, arbitrage, and risk control. The goal of stock investment is to create a portfolio of stocks that maximizes the overall return regarding the risk of stocks in that portfolio (Markowitz, 1952) of which the success heavily depends on the right stock selection. Although as the efficient market hypothesis (EMH) (Fama, 1970) states that in a perfect efficient market, stock markets reflect all available information and follow random pathways, declaring that they cannot be predicted, many studies have shown that the stock market does not conform to the EMH, and the fluctuation of stock is not random walk (Carhart, 1997; Jaworski & Pitera, 2014). Therefore, how to predict the stock trend and establish a profitable portfolio is worth studying. But we cannot deny that the prediction of the market is an extremely difficult task and no strategy can keep winning all the time. From our points of view, this is because the stock price not only reflects the market value of the corresponding company, but also a medium of game for investors. Hence there cannot be a strictly dominated strategy for this game.

With the development of network technology, an ocean of data collected from various sources including the historical trading data, company relationship data, public opinion data and etc. are easier to obtain. Many novel methods using artificial intelligence and machine learning also provide us with the ability to process these types of data. Therefore, using new algorithms and different types of data to assist the stock investment has been an attractive topic in both academia and business field. Some examples are shown in Section 2.

However, three universal defects limit the ability of this data intelligence:

(1) Model evaluation index does not match the application. Most models were built as a classification (on price movement direction) or a regression (on price value) task to predict the change of an index or a stock. But in real investment, the objective is to create a portfolio, not just predict the price of stocks, which means the model needs to adapt to a large range of different type of companies' stock and find those can bring benefit, not predict the future price of stocks.

(2) Missing the screening function. For fund managers to choose stocks, they can selectively filter out stocks that they do not familiar with or they think is risky, and create a portfolio among those they are familiar with. This screening ability can make their portfolio more reliable. However, most methods lack of this screening ability, which will make the established model mediocre because some anomalous stocks are hard to fit by a unified model.

(3) Lack of information fusion. There are many kinds of information related to stock prediction, such as historical trading data, basic information of a company, the media news about the company, etc.

☆ This work was supported by National Natural Science Foundation of China (61933015).
* Corresponding author.
E-mail addresses: liyanrui@zju.edu.cn (Y. Li), 11818434@zju.edu.cn (K. Fu), 3180101495@zju.edu.cn (Y. Zhao), cjyang999@zju.edu.cn (C. Yang).

https://doi.org/10.1016/j.eswa.2022.116629
Received 2 November 2021; Received in revised form 25 January 2022; Accepted 29 January 2022
Available online 17 February 2022

Although there have been studies to combine a variety of data, few of them discussed the characteristics of different types of data. For example, the historical trading data is more suitable for short-term investment while fundamental analysis for long-term investment and the media news is often event-driven. Therefore, how to fuse these data according to their characteristics still needs further research.

To address the mentioned limits, we propose the method with the following advantages: First, we use the average return of the portfolio as our evaluation indicator to build and test our model with Wilcoxon rank-sum test to evaluate the performance of the selected portfolio statistically. Second, the framework has filtering capability, obtained by fusing the novel discriminative model and the news sentiment analysis model. Third, complete data related to stock, including historical trading data, basic information of company, factor data and the media news text, are considered in this fusion model, and the impact range and characteristic of each type of data are discussed. The novelty of our method is reflected in our specific methods to realize the above mentioned advantages. First, we use the Seq2Seq model and t-merge factor model to extract the information in historical trading data and factor data respectively and fuse them to give the basic ranking. The proposed scoring model shows great improvement compared to other methods, especially when they are combined. Second, the novel discriminative model as well as a media sentiment model based on weighted stock relation graph are used to detect anomaly stocks. The two models are designed for the screening purpose which can extract information from factor data and media news data respectively. Finally, in the model fusion process, the screening model gives reward and punishment to the stock list which is ranked by predictive return of scoring models. This fusion method can increase the stability of the selected portfolio and reduce its risk.

The rest of this paper is organized as follows: The related work are represents in Section 2. In Section 3, we introduce the source of data used in proposed framework, the overall structure of the framework and how we evaluate its performance. In Section 4, we describe the detailed structure of each model. In Section 5, the performance of each model and the fusion model are exhibited and discussed, which is followed by the conclusion in Section 6.

2. Related work

From the technical perspective, research related to stock investment mainly uses the following four methods: factor model, time series analysis method, machine learning and deep learning.

(1) Multi-factor model: multi-factor model is a financial model that employs multiple factors in its calculations to explain the stock price, which is widely used for decades. One of the most famous one is the Fama–French Three-factor model, which uses three factors, the market risk, the outperformance of small versus big companies and the outperformance of high $book/market$ versus low $book/market$ companies to explain the USA stock market (FAMA & FRENCH, 1992). Later, it was extended to a five-factor model, added a further two factors: profitability and investment (Fama & French, 2017). Although this model was validated in the USA stock market, it is shown unable to offer a convincing asset pricing model for the stock market in UK (Foye, 2017). Therefore, there is still no widely accepted conclusion on what factors should be used to explain different markets. How to find new factors that can explain the market in a relative validity period is becoming the focus of more and more fund companies. The factors are roughly classified as quality factor, barra factor, risk factor, technical factor, basic factor, etc (Harvey et al., 2015), and more and more factors have been developed. Besides, automatic feature construction are often used to solve the inconvenience of manual feature construction (LeDell & Poirier, 2020; Li & Yang, 2021).

(2) Machine learning: comparing to multi-factor model, machine learning often uses the financial time series data to forecast price of stocks, as a regression task, or to predict the future market trend, as a classification task (Ayala et al., 2021; Bustos & Pomares-Quimbaya, 2020). Several algorithms, like Support Vector Regression (SVR) (Chen & Hao, 2017), Bayesian analysis (Tsay & Ando, 2012),Neural network (Site et al., 2019) and K-Nearest Neighbors (Vijh et al., 2020) have been developed and applied. In Patel et al. (2015a), the authors compare four prediction models, Artificial Neural Network (ANN), Support Vector Machine (SVM), random forest and naive-Bayes, with two approaches for input, and show the performance improve when trend deterministic data are used. In Patel et al. (2015b), the authors proposed two stage fusion approach SVR-ANN, SVR-random forest (RF) and SVR–SVR to predict future market index. Compared to the traditional time series processing methods based on statistical methods such as Autoregressive Integrated Moving Average mode (ARIMA) (Healy, 1964), machine learning methods have better performance for nonlinear multidimensional financial data (Kumar et al., 2021). Besides, some researchers address stock selection problems by clustering to control the investment management (Han & Ge, 2020; Sun et al., 2021; Wang, 2011). In Majumdar and Laha (2020), the authors utilize the method of topological data analysis(TDA) to perform time series classification and discern different stock indexes.

(3) Deep learning: deep learning, also called the deep neural network, is widely studied in finance these days, since it achieved excellent development in natural language processing (NLP) and image processing fields. Its ability to extract features from a large set of raw data without relying on prior knowledge of the predictors model and extract complex hidden patterns from finance data in both temporal and spatial dimensions make the application of deep learning in the capital market increase significantly recently. Most of the deep learning research in the finance field focuses on modeling financial time series using recurrent neural network(RNN) and convolutional neural network (CNN) (Bukhari et al., 2020; Chen et al., 2018; Saud & Shakya, 2020). Besides, considering the different stocks are related by shareholders, industrial chain and etc, this kind of graph structure relationship is introduced into the model by graph convolutional neural network. For example, Chen et al. (2021) proposed the graph convolutional feature based convolutional neural network (GC-CNN) model, considering both stock market information and individual stock information.

(4) Financial text model: text, which contains the most abundant intuitive information, can be automatically analyzed with the development of technology. And a lot of economic news websites provide researchers with sufficient data to study (Li et al., 2018). One of the main research directions is the sentiment analysis of financial text (Seong & Nam, 2021). These works often use machine learning (Wu et al., 2014) and deep learning method (Kilimci & Duvar, 2020; Ren et al., 2020) to extract emotion score from news (Ftiti et al., 2021; Nerger et al., 2021) or social media (Urolagin, 2017) and then valid their correlations to stock price. Mehta et al. (2021) compared the five different machine learning and deep learning methods, including SVM, MNB classifier, linear regression, Naive Bayes and Long Short-Term Memory(LSTM) and LSTM shows the best accuracy. Besides, Gite et al. (2021) applied Explainable AI(XAI) tool on sentimental analysis LSTM model to give an understandable explanation for prediction.

Although forementioned studies have been experimented in the market respectively, fusion of different algorithms and data from various sources still needs further study. For example, Barak et al. (2017) focus on the fusion of different tree-based methods and Li et al. (2021) applied a two-level information fusion approach to examine the effects of peer engagement on social media on stock price synchronicity and compare the effects between epidemic and non-epidemic contexts. However, neither of them considered the screening of stocks, and the data type relevant to stocks is insufficient.

3. Preliminaries

Before describing the details of each model, first of all, the data used in our research, the idea of the overall framework and the way evaluate our models are introduced as shown below.

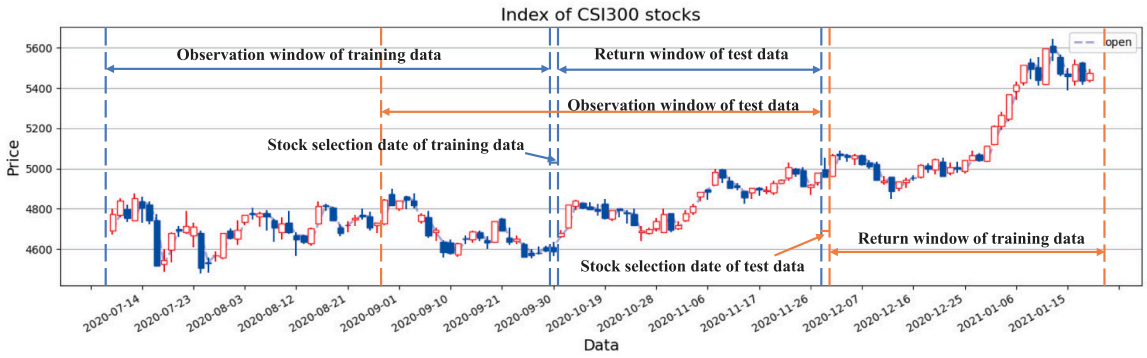

Fig. 1. The visualization of partition of dataset on time series of the CSI 300 index.

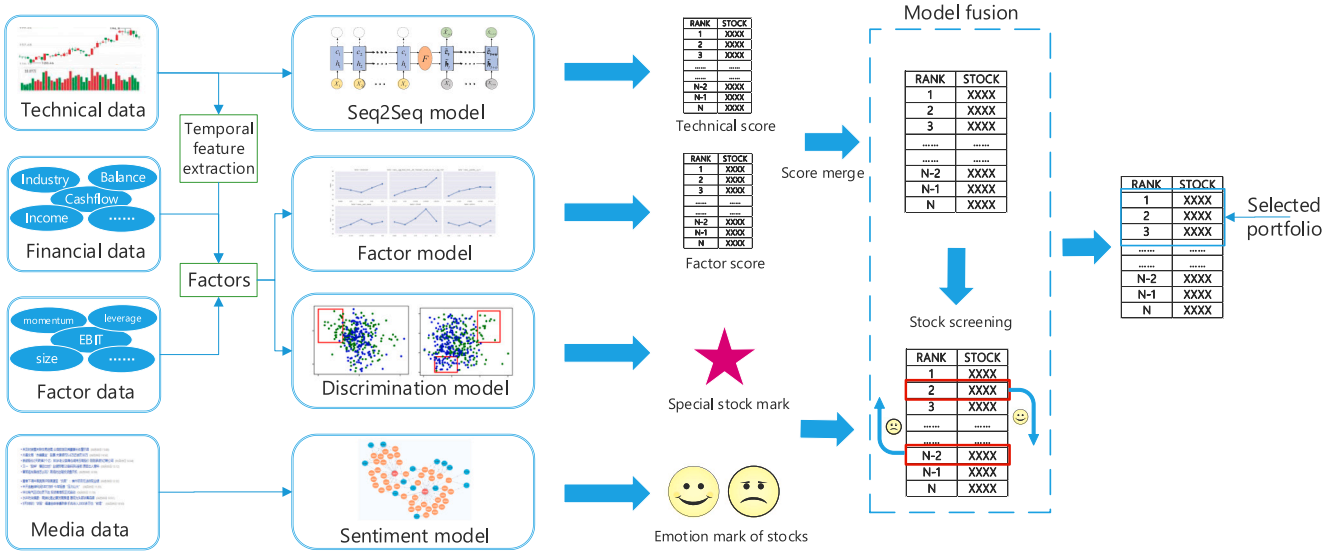

Fig. 2. Schematic of the proposed framework for stock selection.

3.1. Data introduction

The data used in this paper are Chinese stock market data from July 2020 to January 2021, total of 3760 stocks,including the following four types of data:

(1) Historical trading data: The historical trading data presents the time-series stock market data including open price, close price, highest price, lowest price, trading volume and turnover of each stock everyday. Note that we use adjusted price to avoid the discontinuity of stock price when stock split. Besides, we use the change rate of each stock's price relative to the close price of the previous day as the input for modeling and feature extraction to eliminate the impact of the different values of stock's prices.

(2) Financial data: financial data include the basic information of the company, such as its industry categories, shareholder and financial statements updated every quarter, including income statement, cash flow statement and balance statement.

(3) Factor data: more than 300 factors are collected from joinquant[1] which contains different types of factors, such as quality factor, barra factor, risk factor and technical factor. Most of the factors are constructed according to published articles (Zura et al., 2015).

(4) Media data: media data include events related to the stocks and the sentiments of people towards the market and stocks. In order to analysis the public opinion fluctuation during the corresponding time period, we obtained over 30000 pieces of news from industry news and company news sections of Sinafinance,[2] an authoritative financial news platform for Chinese and international investors, with 150 to 200 pieces of news per day, and a total of 1773 companies were mentioned in the news.

3.2. Evaluation method

(1) Average return of portfolio

In previous work about stock prediction, most models were built as a classification (on price movement direction) or a regression (on price value) task, which would cause a large discrepancy on the investment revenue. Besides, even a factor model with a negative R^2 may also yield profitable returns in practice as long as it has the ability to correctly rank the stocks by future return (Feng et al., 2018). Therefore, in our work, the comprehensive return(C-return) and pure return(P-return) are used as the evaluation index of the model, which directly related to the benefits by using the model and are defined as following equation:

$$C - return(D) = \frac{\sum_{t=1}^{D} p_{t_0+t}\,(1-\delta)^t}{p_{t_0}} \tag{1}$$

[1] https://www.joinquant.com/help/api/help#JQData.

[2] https://finance.sina.com.cn/.

$$P - return(D) = \frac{p_{t_0+D}}{p_{t_0}} \tag{2}$$

The two equations above are the index to denote the return of single stock, where $p^i_{t_0}$ is the price of the stock i of the last observation day, and $p^i_{t_0+t}$ is the price of the tth prediction day, with a total of D days. the δ is the depreciation rate we set 0 for the convenience of calculation. The stock selection model will select the K stocks with highest expected return and the $C-return(K, D)$ and $P-return(K, D)$ are used to evaluate the result of the selected set:

$$C - return(K, D) = \frac{\sum_{i\ in\ S} C - return(D)_i}{K} \tag{3}$$

$$P - return(K, D) = \frac{\sum_{i\ in\ S} P - return(D)_i}{K} \tag{4}$$

where S is the stock set of K highest expected return and K is the stock selection scale.

In all model training stages, we set the prediction time D as 36 and use $C - return(36)$ as the label for each stock. The training and testing dataset are split as Fig. 1, models are trained and validated with data from July to August, with the information in observation window as input and stock return in prediction window as label, and finally tested on the data from September 2020 to January 2021. Note the prediction period of the test dataset is completely unknown for the models.

(2) Adjust R squared

In statistic R-square is the proportion of the variation in the dependent variable that is predictable from the independent variable, and often used to evaluate the models' prediction ability. The calculation is as follows:

$$SS_{res} = \sum_i (y_i - f_i)^2 \tag{5}$$

$$SS_{tot} = \sum_i (y_i - 1)^2 \tag{6}$$

$$adjust - R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{7}$$

The difference between the adjust R-square we used and normal R-square is that we use a constant number 1 instead of average return rate to calculate SS_{tot}. That is because the average return rate is a future variable which will cause bias to the evaluation.

(3) Wilcoxon rank-sum test

Wilcoxon rank-sum test, also called Mann–Whitney U test is used in our experiment to evaluate whether the return of selected stocks is greater than the average.

3.3. Structure of framework

As shown in Fig. 2, four models are built to process different types of data. The fusion model includes two parts: the scoring part and screening part. In scoring part, theSeq2Seq model and t-merge factor model extract the information from historical trading data and factor data respectively. In screening part the novel discriminative model as well as a media sentiment model based on weighted stock relation graph are used to detect anomaly stocks from factor data and media news data respectively. Finally, in model fusion process, the screening model give reward and punishment to the stock list which is ranked by predictive return of scoring models. This fusion model with screening function can increase the stability of selected portfolio and reduce its risk. Note the factors in figure includes not only the factor data downloaded, but also features extracted from trading time series using tsfresh (Christ et al., 2016).

4. Model detail

The four models are described in detail in this section.

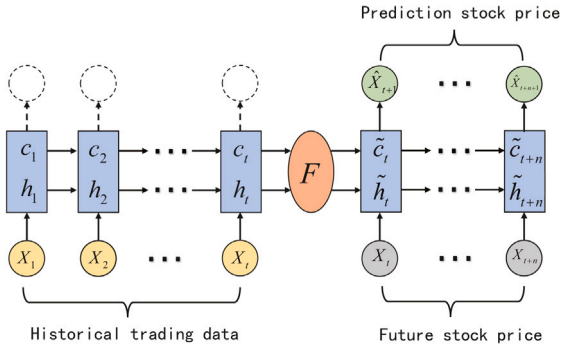

Fig. 3. The structure of the Seq2Seq model.

4.1. Scoring models

The scoring models can give each stock a relatively fair score and build the base rank.

4.1.1. Seq2Seq model

The Seq2Seq model is devised for the historical trading data. As shown in Fig. 3, the Seq2Seq has the encoder–decoder structure. We use the historical trading data in the observation period as the input to predict the stock price sequence in the prediction period. The Seq2Seq network runs as follows. First, the encoder of the Seq2Seq network extracts the vector containing temporal information of the input serialized trading data. Secondly, this vector is put into the decoder part of the network to generate the first prediction of stock price movement. Third, the decoder updates with the prediction return of the previous day as the input and generates the prediction of the next day. The advantage of this network is its ability of fitting complex temporal data and predicting the future return of the stock every day, rather than offering one comprehensive return.

4.1.2. Factor model

We build the factor model based on the proposed t-merge discretization method and histogram-based gradient boosting regression Tree (Ke et al., 2017) with the input variables containing financial information, common factors and temporal features extracted from historical trading data. The whole process includes the following steps:

(1) Data preparation: the factor data consists of two parts, the normal factors downloaded directly from the Joinquant, and the time series factors extracted from the close price and volume by tsfresh (Christ et al., 2016). A total of 947 factors were passed into the subsequent processing.

(2) Factor selection: we use MIC (Albanese et al., 2018; Reshef et al., 2011) to select factors that are highly correlated with the comprehensive return of stocks for its ability to detect various kinds of relationships.

(3) Data discretization: the factor data are discretized by the t-merge algorithm. The t-merge algorithm has the same bottom-up framework as chimerge (Kerber, 1992), but is based on the t value to test the continuous variable for each pair of adjacent intervals. The discretization algorithm is shown in Algorithm 1. Two parameters affect this algorithm: the start bins n_1 determine the computing time and the sensitivity of the algorithm, and the final bins n_2 influence the subsequent modeling result. The time complexity of this algorithm is $O(nlog(n)klog(k))$, where n is the sample number and k is the number of start bins The discretization method can discretize the data with the least loss of information, balancing the stability and sensitivity.

(4) Modeling and validation: As shown in Fig. 4, after discretization, the average return shows great correlation with the value of factor. Some relationships are linear, such as sales to price ratio, the bigger

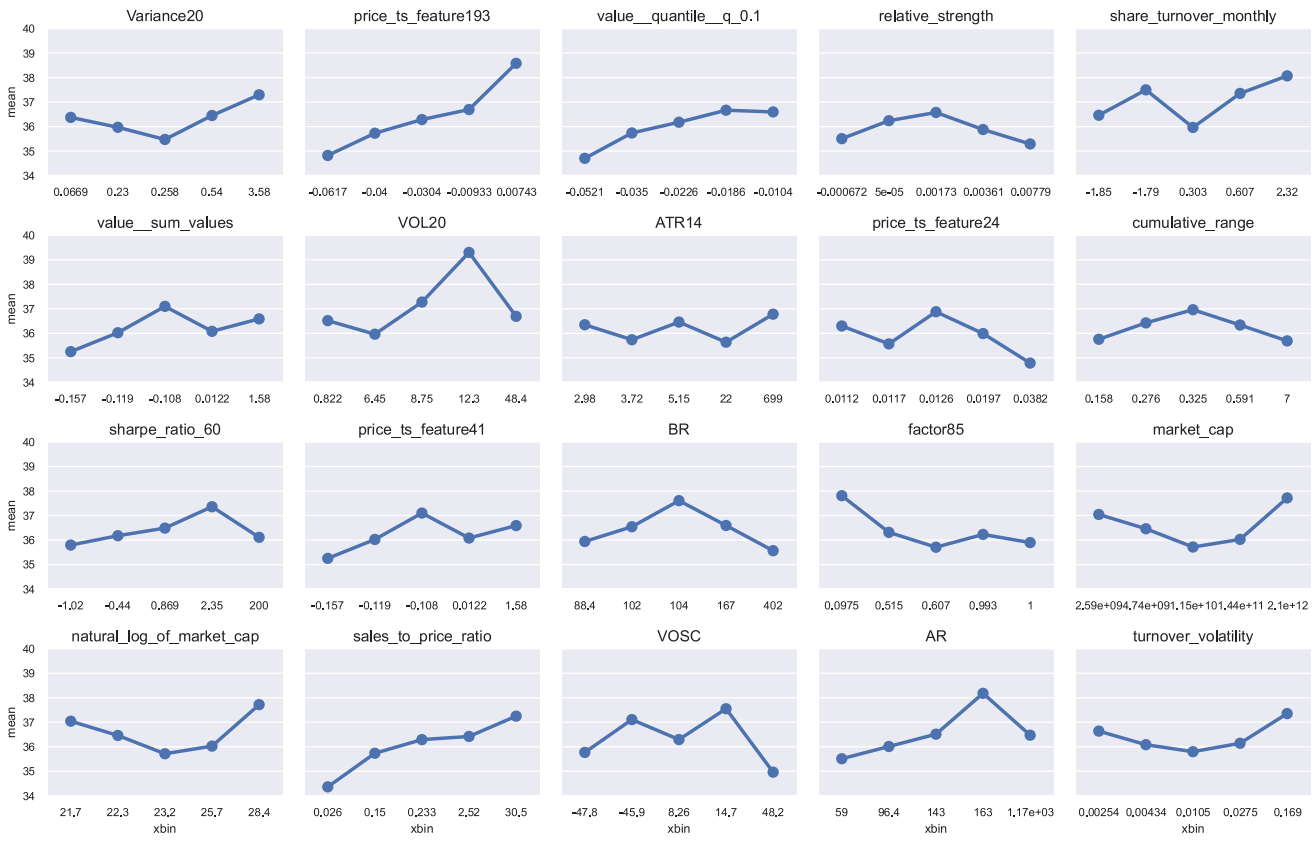

Fig. 4. The relationship between factors and the C-return(after the value of factor is discretized).

Algorithm 1: T-merge discretization algorithm

Data: the continuous value of variables $X(x_1, x_2, \cdots, x_m)$, the corresponding label $Y(y_1, y_2, \cdots, y_m)$
Input: Start bins n_1, Final bins n_2
Output: Discretized variables X_d

$D(d_1, d_2, \cdots, d_{n_1}) \leftarrow$ discretize X into equal-sized buckets based on sample quantiles;
$T(t_1, t_2, \cdots, t_{n_1-1}) \leftarrow$ compute the t value for each pair of adjacent intervals;
while $len(T) > n_2 - 1$ **do**
  $i \leftarrow argmin(T)$;
  delete t_i in T;
  $d_i \leftarrow d_i + d_{i+1}$ (merge the discretization);
  renew the t_{i-1} and t_i;
end
return D

that value, the higher the expected return. However, most relationships are nonlinear, which means the trend of expected return is uncertain, such as market cap, with the increase of the value the expected return may decline first and then rise. Therefore, we use GBDT to build the model because it has the powerful ability to deal with nonlinear relationships.

4.1.3. Comparison

The two proposed scoring model is compared with some classical machine learning model and some new proposed algorithms, as shown in . The comparisons include some classical machine learning methods including Support Vector Regression (SVR), GBDT and Principal Component Analysis-SVM (PCA-SVM); state of art methods including Long Short-Term Memory (LSTM) network and Natural visibility encoding (NVE) and a fusion model SVR+LSTM. Note that the LSTM, Seq2Seq and NVE models are built on historical trading data while others are built on factor data. The label is the return rate of C-return(36) of each stock.

The combination of Seq2Seq model and t-merge factor achieve best accuracy in all results, which has the R^2 greater than 0.15. In all single model, the t-merge factor model has the best result, which is largely improved compare to the direct factor model. This result indicates the proposed t-merge discretization can make fcator model more robust and improve the accuracy. Compared with LSTM, seq2seq can predict every day in the future and show the better result.

4.2. Screening model

The screening models give fusion model the ability to screen the anomalous stocks and reduce the risk of the elected portfolio.

4.2.1. Discrimination model

The factor model does not make full use of the information of all factors because it only selects a small part of factors to model and has to abandon the sensitivity to balance the model's robustness. In fact, more information can be mined from factor data if we focus on the anomalous stocks that will surge or drop rapidly in the future. In this section, we will introduce a novel method to find simple rules to measure the stocks from large dimension factor data. To sum up, the two proposed methods have shown great improvement compared with their respective comparison methods, and the two methods can complement each other and achieve the best performance by using data with different characteristics.

The discrimination model is an ensemble model which consists of many base models. In Fig. 5, we display four base models to explain how this method works. The base model only takes two dimensions

Algorithm 2: Discrimination model

Data: Stock set S, the factors of stocks $X_1, X_2, \cdots, X_n$(training data), the C-return of stocks(training data) Y
Input: the factor of stock $x_1, x_2, \cdots, x_n$(test data), base model number m, distance number k, threshold h
Output: the discrimination model, the predict return of the test stock $\hat{y}$

According to the return Y, the most profitable 300 stocks are selected into S_+ and the 300 stocks with the worst return are selected into S_- ;
for $i = 1;\ i < n;\ i++$ **do**
  for $j = i;\ j < n;\ j++$ **do**
    $M[i,j] \leftarrow$ compute MIC between X_i and X_j;
    $M[i,j]^+ \leftarrow$ compute MIC between $X_i[S_+]$ and $X_j[S_+]$;
    $M[i,j]^- \leftarrow$ compute MIC between $X_i[S_-]$ and $X_j[S_-]$;
    $M_{diff}[i,j] \leftarrow max(M[i,j], M[i,j]^+, M[i,j]^-) - min(M[i,j], M[i,j]^+, M[i,j]^-)$;
  end
end
$base\ model \leftarrow$ select the m pairs of X_i and X_j with the largest $M_{diff}[i,j]$;
return the base model;
prediction;
for i, j *in each base model* **do**
  $d_+ \leftarrow$ sum of the distance of the nearest k samples in S_+;
  $d_- \leftarrow$ sum of the distance of the nearest k samples in S_-;
  if $|log(d_+/d_-)| > h$ **then**
    $discriminative\ score = discriminative\ score + log(d_+/d_-)$;
  end
end
return discrimination score

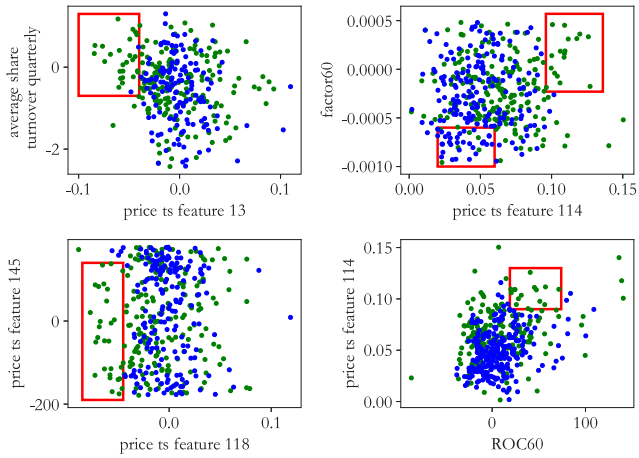

Fig. 5. The idea of discrimination model, for each diagram (representing each base model), the coordinate axis are two selected factors and the points in graph represent stocks with huge rise or fall (green and blue respectively). The boxed areas are scoring areas.

of factors as input. The samples are separated into two classes(green for surge stocks, blue for drop stocks). Although there is no difference between surge stocks and drop stocks overall, when we focus on the local areas, we will find the distribution of stock types is different (as boxed in red). These are the scoring areas that can screen the discriminative stocks. Each base model can only screen out a few samples of stocks, but through the ensemble of many base models, a large number of stocks that have the characteristics to rise or fall sharply in the future can be screened out. The main problem is how to find the base model containing the score area, and how to find the score area in the base model. The pseudocode of the model construction programme is shown in Algorithm 2. In a word, we use the difference of MIC among the subsets of stocks to find the pairs of factors containing the scoring area, and utilize KNN to find whether a sample is in the scoring area. This discrimination model is a type of screening model. The screening model aims at finding a subset of samples rather than estimating all samples which, to some extent, is similar to selecting stocks by qualitative analysis.

4.2.2. Media model

Another screening model we built in this paper is the media emotion model. The media emotion model scores the emotion grade for stocks mentioned in financial news based on the weighted stock relation graph. The construction process of the media emotion model is shown in Fig. 6. First, we use basic stock data and historical trading data to construct weighted stock relationship graphs as shown in Fig. 7. A stock node is connected with the industry nodes and concept nodes it belongs to. The weight of their connection are computed according to the correlation between the nodes between the trend of the stock and the trend of the industry(concept). Second, the emotion mark of each entity mentioned in the news is computed by a deep-learning NLP model provided by ERNIE 2.0 constructed on PaddlePaddle structure (Sun et al., 2019). This model is pretrained on the large Chinese language corpus and we fine-tuned it with up to 1393 manually annotated positive news texts as well as 1393 negative ones. This model can extract entities including stocks, industries and concepts we set as keywords previously and provide an emotion score to each of them respectively. Third the stock emotion mark are summed according to emotion score of itself and the industries and concepts connected to it. Because the media emotion is often event-driven, the impact of public emotion on the stock market is continuous over time. We use an influence curve(Fig. 8) to simulate the impact power of the media emotion as time goes on. We assume that after the release of the news, the influence will decline exponentially over time. Based on this curve, we sum emotion score of each stocks at the current moment over a period of time as the following equation.

$$s = \sum w_c s_c + w_i s_i + s_s \tag{8}$$

The w_c, w_i are weights of linkings. The s_c, s_i are the media scores of the concept and industry that linked to the stock, s_s is the emotion score of the stock itself, and all media scores are calculated on the same period of time.

The novelty of this method include two parts: first, it consider the characteristic of the media emotion that public sentiment is often event-driven and its impact is continuous over time. Besides, in this model, the emotion score of company is not only about itself but also influenced by the public opinion of the industry and concept related to it. These two characteristics of the model make it better to capture public emotions, which makes the ability function of fusion model more effective.

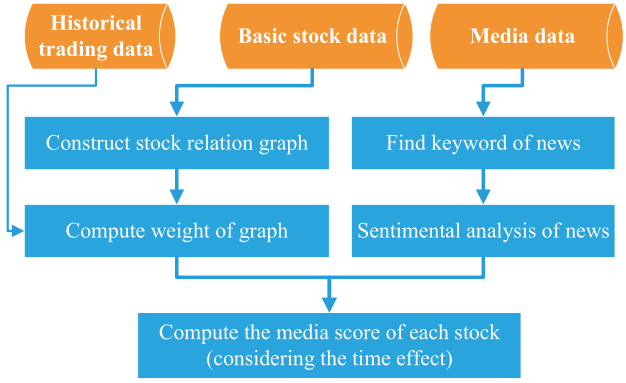

Fig. 6. The construction process of media model.

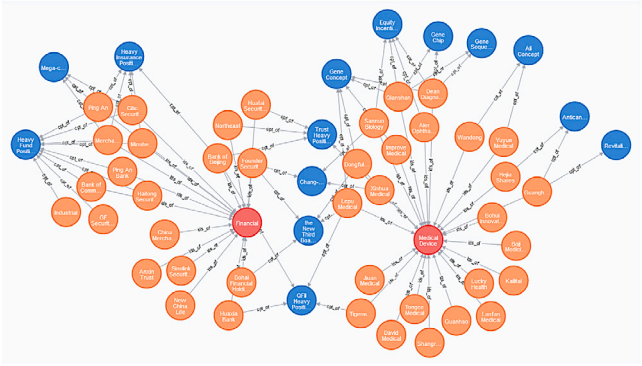

Fig. 7. Stock relationship graph (part). Three types of nodes are in the graph: the orange ones represent stock nodes, the nodes in red are industry nodes and the nodes in blue are concept nodes. Note that the stocks included in industry and concept nodes are defined in data according to the expertise knowledge. Normally, a concept includes some companies with the common ground, which means that they may belong to the same industrial chain (like concept 'gene chip'), or are related to the same companies (like concept 'Ali related').

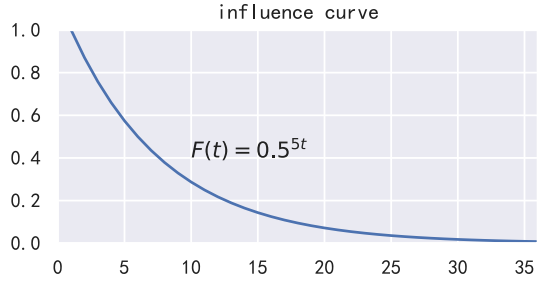

Fig. 8. Influence curve.

Table 1
Prediction performance of different scoring methods.

Algorithm	Rmse	Adjust-r^2
SVR	0.1006	0.0252
LSTM	0.0905	0.0622
GBDT	0.0932	0.0024
LightGBM (Ke et al., 2017)	0.0889	0.0687
T-merge factor	**0.0786**	**0.1374**
LSTM	0.0989	0.0552
Seq2Seq	**0.0868**	**0.0774**
Seq2Seq+t-merge factor	**0.0734**	**0.1528**
NVE (Huang et al., 2021)	0.0822	0.1233
PCA-SVR	0.0934	0.0931
LSTM+GBDT	0.0794	0.1336

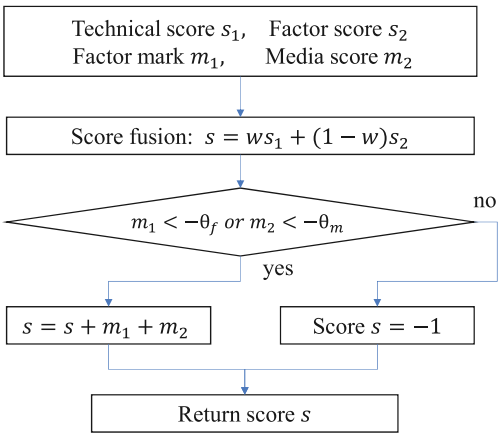

Fig. 9. Schematic diagram of model fusion.

5. Model fusion and result

In the previous chapter, we have introduced the structure and training method of each model, and validated them on the data from July 2020 to November 2020. In this chapter, we will test the trained single models and the fusion model on the test dataset from September 2020 to January 2021.

5.1. Model fusion

The fusion mechanism fuse the screening model and scoring model with different rules as shown in Fig. 9. The two types of information: score calculated by scoring model s_1, s_2 and marks generated by screening models m_1, m_2 are the input. First, two scores s_1, s_2 are weighted averaged to calculate the score s, by which the base ranking generated. Then two screening models will veto the stocks that are marked as bad stocks and give reward to stocks marked as good ones. This fusion process is determined by three parameters: the scoring weight ω and $1-\omega$ for Seq2Seq model and factor model and the θ_f and θ_m, the threshold to determine whether to veto the stocks or not. In our model, we set $\omega = 0.5$, $\theta_f = 1$ and $\theta_m = 10$. Note that this fusion mechanism is designed under the fact that there is no effective short investment in china's A market so it is impossible for the model to predict the fall of stocks to earn benefits.

5.2. Results

Each separate model and the fusion model are verified on the test dataset, and the results are shown in Table 1. We can find almost all models have positive extra return under different trading conditions (K or D), which proves their ability to select the profitable stocks. The fusion model can achieve best return in most trading conditions, especially when $K = 100$ and $K = 1000$. Besides, we can find that the increase of the D lead to the increase of the return, which means the rising trend of the selected stock is enduring.

Next, we will further analyze the result through three charts. The Fig. 10 shows the decline of daily return when K increases, which means the models can correctly rank the future return of the stocks but not a rise-or-fall classification. The Fig. 11 presents the relationship between D and the daily return. The daily returns are higher when D is 36. This shows the models are sensitive to the days of prediction period. We can retrain the model with different labels when predicting in different periods of time. The Fig. 12 shows the relationships of two results evaluation indexes which also means two trading strategies. The abscissa and ordinate of it represent the values of a class of indicators, and each point in the chart represents the corresponding values of the results of the two indicators of one setting.

Table 2
The experiment results with different setting of K and D. Two values, extra return rate(ER) and p-value of Wilcoxon rank-sum test are shown in table to display the performance of the selected portfolio in different settings.

		Seq2Seq		Factor model		Media model		Discriminative model		Fusion model	
		C-R	P-R	C-R	P-R	C-R	P-R	C-R	P-R	C-R	P-R
$K = 100, D = 15$	ER	0.421	1.965	0.088	2.187	0.727	2.248	0.477	2.102	**0.876**	**3.630**
	p-value	0.441	0.077	0.787	***0.039***	0.099	***0.046***	0.226	***0.017***	0.294	***0.018***
$K = 200, D = 15$	ER	0.299	1.114	0.435	1.842	1.202	**2.724**	0.250	0.682	**1.558**	2.496
	p-value	0.557	0.143	0.224	0.068	0.089	***0.039***	0.352	0.282	0.147	***0.036***
$K = 500, D = 15$	ER	0.119	0.257	0.630	1.600	0.626	1.402	−0.109	−0.129	**0.815**	**1.787**
	p-value	0.589	0.443	0.495	0.121	0.222	0.110	0.796	0.855	0.176	0.057
$K = 1000, D = 15$	ER	−0.025	−0.134	0.192	0.619					**0.279**	**0.773**
	p-value	0.774	0.784	0.733	0.521					0.544	0.454
$K = 100, D = 36$	ER	2.329	4.418	2.460	6.883	3.528	6.663	2.969	3.119	**3.793**	**7.924**
	p-value	***0.039***	***0.002***	***0.036***	*<0.001*	***0.008***	***0.001***	***0.033***	***0.019***	***0.018***	*<0.001*
$K = 200, D = 36$	ER	2.173	4.936	2.317	5.847	**3.377**	**6.551**	1.996	2.871	3.002	6.180
	p-value	0.053	***0.002***	***0.046***	***0.001***	***0.020***	***0.001***	***0.037***	***0.024***	***0.021***	***0.001***
$K = 500, D = 36$	ER	1.032	3.146	1.996	**4.114**	**1.934**	3.710	0.384	0.565	1.854	3.610
	p-value	0.133	***0.029***	***0.042***	***0.003***	0.068	***0.016***	0.656	0.677	0.078	***0.013***
$K = 1000, D = 36$	ER	0.397	1.388	0.822	1.642					**1.093**	**1.971**
	p-value	0.433	***0.090***	0.239	***0.049***					0.076	***0.041***
$K = 100, D = 60$	ER	3.303	4.518	4.667	6.769	5.064	5.534	3.089	4.544	**5.647**	**9.229**
	p-value	***0.022***	***0.002***	***0.004***	*<0.001*	***0.002***	***0.002***	***0.024***	***0.003***	***0.002***	*<0.001*
$K = 200, D = 60$	ER	1.378	2.220	4.308	**7.086**	4.268	3.984	2.101	2.906	**4.681**	6.617
	p-value	0.053	***0.015***	***0.005***	*<0.001*	***0.007***	***0.006***	***0.041***	***0.034***	***0.003***	***0.001***
$K = 500, D = 60$	ER	2.076	4.442	3.344	5.273	2.644	2.745	0.168	0.577	**3.883**	**5.443**
	p-value	***0.029***	***0.004***	***0.013***	***0.001***	***0.032***	***0.035***	0.545	0.345	***0.003***	***0.006***
$K = 1000, D = 60$	ER	1.535	2.385	1.648	2.761					**1.649**	**2.939**
	p-value	0.067	***0.049***	0.114	***0.043***					0.088	***0.026***

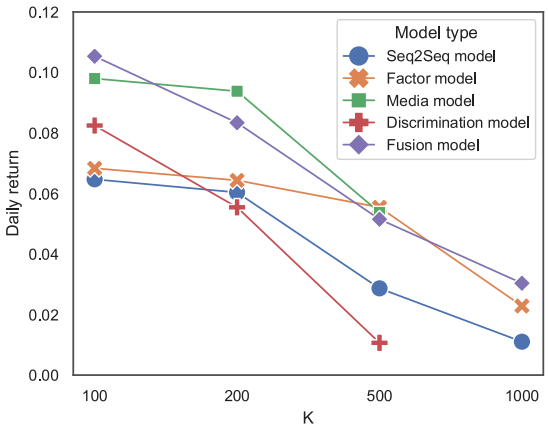

Fig. 10. The change of daily return of models with different selection scale(under time period $D = 36$).

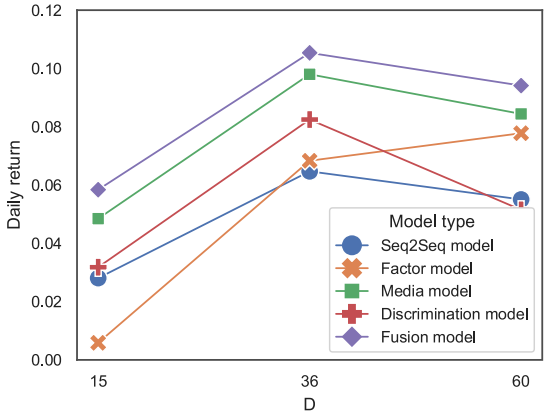

Fig. 11. The change of daily return of models with different prediction time(under selection scale $K = 100$).

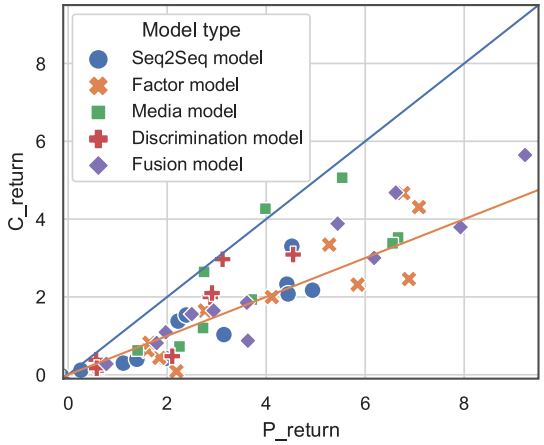

Fig. 12. Relationship of two evaluation indexes.

The ratio of the two types of return can reflect the trend of the stock in the prediction period. If the stock price rises by the same amount every day, the ratio of $P - return$ and $C - return$ will be 2, and if the price rises by the same rate every day, the ratio will be much larger. However, the most of the results are smaller than 2, which indicate that the model can find the stocks that is profitable even if they fall back at the end of the prediction period. Besides, the ratio seems higher when prediction period is short, indicating the stocks may maintain the upward trend in this period.

The result of our fusion model is also compared to other algorithms under the experiment setting $K = 100, D = 36$, which means the portfolio contains 100 stocks and the label of each model is the $C - Return$ of 36 days. For other algorithm, the portfolio is created by selecting the stocks with the top 100 maximum predictive return. The result is shown in Table 3. The fusion model is much better than other models, which is mostly because the two proposed scoring model can

Table 3
The comparison of performance of portfolio selected by different methods.

Algorithm	$C-R\ (p-value)$	$P-R\ (p-value)$
SVR	2.24(0.033)	3.43(0.024)
LSTM	3.27(0.016)	4.89(0.005)
GBDT	1.03(0.178)	4.14(0.014)
Factor	1.52(0.069)	3.33(0.039)
Fusion model	**3.793(0.018)**	**7.924** (<**0.001**)
PCA+SVR	0.797(0.285)	1.219(0.248)
NVE	2.19(0.042)	6.32(0.002)

complement each other and fully use the different types of data and the screening function of the fusion model.

5.3. Discussion

5.3.1. The effectiveness of model

From the Table 2, the factor model and the media model, the most studied and mature models, have best results in four separate models. As shown in Fig. 11, the factor model seems to have more advantages in long time prediction, because it is the only one with daily income rising when $D = 60$. Besides, the daily return of screening models drop rapidly with K increasing. It indicates that this kind of model has a strong polarization, that is, good stocks and bad stocks that screened with high scores are more reliable than ones selected by scoring models, while the stocks screened with low score is relative unreliable. Therefore, we set the threshold of screening model θ in fusion mechanism to eliminate unreliable results in screening model. Due to the fusion mechanism that gives bonus to the good stocks and the veto to the bad stocks, the performance of the fusion model is significantly improved compared to the single model, especially when K is 100 and 1000.

5.3.2. The effectiveness of data

In some people's view, historical data cannot provide any help to the stock forecast, because it is too simple for investors to collect the data, only the financial data commonly used in the qualitative analysis is effective. However, counting the sources of factors with high correlation in the factor model(among top 40 correlated factors 26 are factors extracted from historical trading time series), we find the historical trading data is capable of providing sufficient useful information whereas some factors that have been proved to be effective in many papers are ineffective. This may be due to the characteristics of China's stock market, or the temporally abnormal circumstances of stocks in the post epidemic period.

Besides, the effectiveness of media data has a strong relationship with the source and processing method of media data. In this work, we just build a simple framework to process the media data, however, the result proves to be excellent. Therefore, we think it is very important, and a customized NLP model may lead to a better result.

5.3.3. Model fusion mechanism

When designing the model fusion mechanism, we have completely different rules for the good stocks and bad stocks obtained from the screening model. We give a more severe punishment, the veto, to the bad stocks compared to the award we give to the good ones. Therefore, the fusion model can get rid of selecting bad stocks. This is mainly because there is no effective short mechanism in China's A-share market and the accurate prediction of stocks fall is not as important as the accurate prediction of stocks rise. Besides, although investors are often described as risk averse, many of them are, rather, loss averse. This mechanism encourages fusion model to chase risky stocks that may bring more profits and abandon the ones that will bring more losses.

6. Conclusion

In this paper, we built a framework that can use four types of financial data to rank the stocks in A market of China for investment stock selection. The framework contains 4 models, which can be divided into scoring models and screening models. By fusion of four models, The P-return(100,36) of the final result on the test dataset is 9.2% higher than the average value, 2.5% higher than the maximum of the single model, which is a significant improvement. The limitation of the work is as follow: first, the framework is validated on a relative short time, the robustness of the model and the expire time of the model need be validated more in practice. Second, the calculation time of the model is more than 3 h, so it is only suitable for medium and long-term investment, not for high-frequency trading. Last but not least, how to combine human knowledge with the proposed framework need more exploration.

CRediT authorship contribution statement

Yanrui Li: Methodology, Software, Data curation, Writing – original draft. **Kaiyou Fu:** Data curation, Factor model construction, Writing – review & editing. **Yuchen Zhao:** Data curation, Media model construction, Writing – review & editing. **Chunjie Yang:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by National Natural Science Foundation of China (61933015).

References

Albanese, D., Riccadonna, S., Donati, C., & Franceschi, P. (2018). A practical tool for maximal information coefficient analysis. *GigaScience*, *7*(4), giy032.
Ayala, J., Garca-Torres, M., Noguera, J. L. V., Gmez-Vela, F., & Divina, F. (2021). Technical analysis strategy optimization using a machine learning approach in stock market indices. *Knowledge-Based Systems*, *225*, Article 107119.
Barak, S., Arjmand, A., & Ortobelli, S. (2017). Fusion of multiple diverse predictors in stock market. *Information Fusion*, *36*, 90–102.
Bukhari, A. H., Raja, M. A. Z., Sulaiman, M., Islam, S., Shoaib, M., & Kumam, P. (2020). Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting. *IEEE Access*, *8*, 71326–71338.
Bustos, O., & Pomares-Quimbaya, A. (2020). Stock market movement forecast: A systematic review. *Expert Systems With Applications*, *156*, Article 113464.
Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal Of Finance*, *52*(1), 57–82.
Chen, Y., & Hao, Y. (2017). A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. *Expert Systems With Applications*, *80*, 340–355.
Chen, W., Jiang, M., Zhang, W.-G., & Chen, Z. (2021). A novel graph convolutional feature based convolutional neural network for stock trend prediction. *Information Sciences*, *556*, 67–94.
Chen, W., Yeo, C. K., Lau, C. T., & Lee, B. S. (2018). Leveraging social media news to predict stock index movement using RNN-boost. *Data & Knowledge Engineering*, *118*, 14–24.
Christ, M., Kempa-Liehr, A. W., & Feindt, M. (2016). Distributed and parallel time series feature extraction for industrial big data applications. ArXiv E-Prints, pages, arXiv:1610.07717.
Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal Of Finance*, *25*(2), 383–417.
FAMA, E. F., & FRENCH, K. R. (1992). The cross-section of expected stock returns. *The Journal Of Finance*, *47*(2), 427–465.
Fama, E. F., & French, K. R. (2017). International tests of a five-factor asset pricing model. *Journal Of Financial Economics*, *123*(3), 441–463.
Feng, F., He, X., Wang, X., Luo, C., Liu, Y., & Chua, T. (2018). Temporal relational ranking for stock prediction. CoRR, abs/1809.09441.

Foye, J. (2017). Testing alternative versions of the fama-french five-factor model in the UK. *SSRN Electronic Journal*.

Ftiti, Z., Ben Ameur, H., & Louhichi, W. (2021). Does non-fundamental news related to COVID-19 matter for stock returns? evidence from shanghai stock market. *Economic Modelling*, *99*, Article 105484.

Gite, S., Khatavkar, H., Kotecha, K., Srivastava, S., Maheshwari, P., & Pandey, N. (2021). Explainable stock prices prediction from financial news articles using sentiment analysis. *PeerJ Computer Science*, *7*, Article e340.

Han, J., & Ge, Z. (2020). Effect of dimensionality reduction on stock selection with cluster analysis in different market situations. *Expert Systems With Applications*, *147*, Article 113226.

Harvey, C. R., Liu, Y., & Zhu, H. (2015). And the cross-section of expected returns. *The Review Of Financial Studies*, *29*(1), 5–68.

Healy, M. J. R. (1964). Smoothing, forecasting and prediction of discrete time series. *Journal Of The Royal Statistical Society: Series A (General)*, *127*(2), 292–293.

Huang, Y., Mao, X., & Deng, Y. (2021). Natural visibility encoding for time series and its application in stock trend prediction. *Knowledge-Based Systems*, *232*, Article 107478.

Jaworski, P., & Pitera, M. (2014). On spatial contagion and multivariate GARCH models. *Applied Stochastic Models In Business And Industry*, *30*(3), 303–327.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *30*, *Advances in neural information processing systems*. Curran Associates, Inc..

Kerber, R. (1992). ChiMerge: Discretization of numeric attributes. In *Proceedings of the 10th national conference on artificial intelligence. San Jose, CA, July 12-16, 1992.*.

Kilimci, Z. H., & Duvar, R. (2020). An efficient word embedding and deep learning based model to forecast the direction of stock exchange market using Twitter and financial news sites: A case of Istanbul stock exchange (BIST 100). *IEEE Access*, *8*, 188186–188198.

Kumar, D., Sarangi, P. K., & Verma, R. (2021). A systematic review of stock market prediction using machine learning and statistical techniques. *Materials Today: Proceedings*.

LeDell, E., & Poirier, S. (2020). H2O autoML: Scalable automatic machine learning. In *7th ICML workshop on automated machine learning (AutoML)*.

Li, Q., Chen, Y., Wang, J., Chen, Y., & Chen, H. (2018). Web media and stock markets : A survey and future directions from a big data perspective. *IEEE Transactions On Knowledge And Data Engineering*, *30*(2), 381–399.

Li, Y., & Yang, C. (2021). Domain knowledge based explainable feature construction method and its application in ironmaking process. *Engineering Applications Of Artificial Intelligence*, *100*, Article 104197.

Li, L., Zhu, F., Sun, H., Hu, Y., Yang, Y., & Jin, D. (2021). Multi-source information fusion and deep-learning-based characteristics measurement for exploring the effects of peer engagement on stock price synchronicity. *Information Fusion*, *69*, 1–21.

Majumdar, S., & Laha, A. K. (2020). Clustering and classification of time series using topological data analysis with applications to finance. *Expert Systems With Applications*, *162*, Article 113868.

Markowitz, H. (1952). PORTFOLIO selection*. *The Journal Of Finance*, *7*(1), 77–91.

Mehta, P., Pandya, S., & Kotecha, K. (2021). Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Computer Science*, *7*, Article e476.

Nerger, G.-L., Huynh, T. L. D., & Wang, M. (2021). Which industries benefited from Trump environmental policy news? evidence from industrial stock market reactions. *Research In International Business And Finance*, *57*, Article 101418.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015a). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems With Applications*, *42*(1), 259–268.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015b). Predicting stock market index using fusion of machine learning techniques. *Expert Systems With Applications*, *42*(4), 2162–2172.

Ren, Y., Liao, F., & Gong, Y. (2020). Impact of news on the trend of stock price change: an analysis based on the deep bidirectiona LSTM model. *Procedia Computer Science*, *174*, 128–140, 2019 International Conference on Identification, Information and Knowledge in the Internet of Things.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, *334*(6062), 1518–1524.

Saud, A. S., & Shakya, S. (2020). Analysis of look back period for stock price prediction with RNN variants: A case study on banking sector of NEPSE. *Procedia Computer Science*, *167*, 788–798, International Conference on Computational Intelligence and Data Science.

Seong, N., & Nam, K. (2021). Predicting stock movements based on financial news with segmentation. *Expert Systems With Applications*, *164*(November 2018), Article 113988.

Site, A., Birant, D., & I0603k, Z. (2019). Stock market forecasting using machine learning models. In *2019 innovations in intelligent systems and applications conference* (pp. 1–6).

Sun, L., Wang, K., Balezentis, T., Streimikiene, D., & Zhang, C. (2021). Extreme point bias compensation: A similarity method of functional clustering and its application to the stock market. *Expert Systems With Applications*, *164*, Article 113949.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2019). ERNIE 2.0: A continual pre-training framework for language understanding. arXiv preprint arXiv:1907.12412.

Tsay, R. S., & Ando, T. (2012). BayesIan panel data analysis for exploring the impact of subprime financial crisis on the US stock market. *Computational Statistics & Data Analysis*, *56*(11), 3345–3365, 1st issue of the Annals of Computational and Financial Econometrics Sixth Special Issue on Computational Econometrics.

Urolagin, S. (2017). Text mining of tweet for sentiment classification and association with stock prices. In *2017 international conference on computer and applications* (pp. 384–388).

Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia Computer Science*, *167*, 599–606, International Conference on Computational Intelligence and Data Science.

Wang, R. (2011). Stock selection based on data clustering method. In *2011 seventh international conference on computational intelligence and security* (pp. 1542–1545).

Wu, D. D., Zheng, L., & Olson, D. L. (2014). A decision support approach for online stock forum sentiment analysis. *IEEE Transactions On Systems, Man, And Cybernetics: Systems*, *44*(8), 1077–1087.

Zura, K., Geoffrey, L., & Igor, T. (2015). 101 Formulaic alphas. *Ssrn Electronic Journal*.