Information Systems and Economics (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2024.050304
ISSN 2523-6407 Vol. 5 Num. 3

Research on the Influencing Factors and Estimation Model of Enterprise Value: Evidence from China's New OTC Market

Shuxia Wang*, Han Liu, Yang Yang, Chang Liu, Yuqi Zhao, Canjie Zheng

Beijing Institute of Petrochemical Technology, Beijing, 102617, China
**Corresponding author*

Keywords: China's new OTC market; Enterprise valuation model; Random forest; XGboost; Ensemble learning

Abstract: As a national securities trading venue for small and medium-sized enterprises, China's new over-the-counter (OTC) market, also known as the "new third board", plays an important role in China's economic development. To accurately estimate the value of enterprises listed on the China's new OTC market and help investor to make better investment decisions, we first construct a comprehensive factor system that may affect the market value of enterprises. Then, we design an ensemble valuation model based on XGboost and random forest with the selected influencing factors. Experimental results show that the proposed valuation model can significantly improve the accuracy of the valuation for the enterprises listed on China's new OTC market compared with traditional valuation models.

1. Introduction

As a national securities trading venue for small and medium-sized enterprises, China's new over-the-counter (OTC) market, also known as the "new third board", plays an important role in China's economic development. Since the establishment of China's new OTC market in 2006, it has shown a momentum of vigorous development, and has gradually become one of the trading venues with the largest number of listed enterprises in the world. However, at the same time, the problems such as difficulty in pricing and chaotic valuations of enterprises listed on the new OTC market are also emerging. These have severely affected the development of the new OTC market.

Compared with the enterprises listed on the main stock exchange, the enterprises listed on the new OTC market have the following main characteristics: (1) most of them are in the initial stage of development, with unsatisfactory financial situation, relatively low income or even loss; (2) the development of enterprises may be influenced by a lot of uncertain factors, with large fluctuation of performance and high investment risk; and (3) most of them belong to high-tech industries, with strong innovation, less comparable enterprises and lack of reference indicators for valuation.

Due to the characteristics of enterprises listed on the new OTC market, traditional valuation theories and models, such as MM theory [12], CAPM model [14] and B-S option pricing model [3], will lead to large deviation of the valuation results. In the new OTC market, one side is the

high-tech enterprises which are in urgent need of financing, while the other side is the investors who cannot choose the right investment targets since lack of reasonable and effective valuation methods. Therefore, there is an urgent need for a reasonable valuation method for both ofthe two sides.

Most of existing research focused on China's A-share market. With the rapid development of the new OTC market, several scholars tried to use traditional valuation methods, such as discounted cash flow method, economic value added method and real option method, to study the valuation problem for enterprises listed on the new OTC market [5][9][13][17]. However, these methods rely on a lot of manual setting parameters. Therefore, in their empirical analysis, only a few enterprises are selected for demonstration, and the generalization of these methods needs to be strengthened.

To avoid heavy manual setting of parameters, some scholars tried to machine learning based models to study the problem of enterprise valuation[1][2][4][6][7][8][10][11][15][16]. However, they focus on the linear models, such as multiple linear regression models, with the assumption that the selected factors linearly affect the enterprises' value[1][8][17], while ignoring their nonlinear interaction. In addition, they only make use of the technical indicators, such as corporate fundamental financial factors and turnover rates[2][7][10], while ignoring other important factors like macroeconomics, market conditions, industry, external influences and internal factors of enterprises.

In recent years, the development of machine learning and artificial intelligence has provided new ideas for the valuation of enterprises listed on the new OTC market. Especially, the extreme gradient boosting tree (XGBoost) algorithm has a wide range of applications and is universal for discrete or continuous data. The random forest (RF) algorithm has been proved to perform well on multi-dimensional data. In this paper, XGBoost and RF algorithms are used as the base models for enterprise valuation, based on which we propose an ensemble model to get better valuation performance.

The rest of this paper is organized as follows. First, we build a comprehensive system of factors, which may affect the market value of the new OTC market listed enterprises, in Section 2. Second, we propose an XGBoost-RF fusion model based on stacking integration to study the valuation of enterprises listed on the new OTC market in Section 3. Then, we evaluate the proposed model and compare it with several baseline models in Section 4. Last, we conclude this paper in Section 5.

2. Construction of Influencing Factor System

Enterprises listed on the new OTC market are different from those on the A-shared market in terms of development stage, company scale, financing channels, and business models. The market value of new OTC enterprises is easily affected by various factors such as macroeconomics, market conditions, industry characteristics, individual stock fundamentals, technical strength, and governance efficiency. To construct a comprehensive and effective factor system for the valuation model of new OTC enterprises, we construct and select the factors according to the following standards.

(1) Comprehensiveness: the factor system should cover all kinds of factors from macro to individual, from external to internal.

(2) Comparability: the data of influencing factors of the same enterprises in different periods should be comparable, and the factor statistics between different enterprises should be consistent.

(3) Adaptability: the selected factors should be suitable for characteristics of enterprises listed on the new.

(4) Dynamic: the selected factors should be able to reflect the latest economic and industrial development trends and market changes, and reflect the future growth potential of new OTC market

listed enterprises.

Based on previous research experience, combined with the characteristics of the new OTC market, and taking into account comprehensiveness, comparability, adaptability, and dynamic standards, we select 67 market value influencing factors in 13 categories, as shown in Table 1.

Table 1: Influencing factor system for the valuation of enterprises listed on new OTC market

Categories	Influencing Factors	Categories	Influencing Factors
Profitability	Gross Profit Margin	Macroeconomic	Year-on-year GDP Growth Rate
	Net Profit Margin		Year-on-year CPI Growth Rate
	ROE		Year-on-year M2Growth Rate
	Return on Invested Capital		SHIBOR 6 Months
Enterprise Scale	Total Assets	Market Quotation	Average Daily Turnover Rate of Growth Enterprises Market
	Total Equity		Average Daily Turnover Rate of Small and Medium Enterprise Board
	Net Profit		
	Operating Revenue		
	Registered Capital		
Operating and Debt Paying Ability	Debt to Asset Ratio		Average Daily Turnover Rate of CSI 300 Section
	Current Ratio		Average Daily Turnover Rate of New OTC Market Section
	Acid-test Ratio		Average PE of Growth Enterprises Market
	Receivables Turnover Ratio		Average PE of Small and Medium Enterprise Board
	Total Assets Turnover		Average PE of CSI 300
	Fixed Asset Turnover		Average PE of New OTC Market
Growth Opportunity	Year-on-year Operating Revenue Growth Rate	Industry	Proportion of Transaction Volume in Investment-oriented Quarternary Industries
	Year-on-year Total Assets Growth Rate		Proportion of Gross Revenue in Investment-oriented Quarternary Industries
	Year-on-year Net Profit Growth Rate		Transaction Volume of Investment-oriented Tertiary Industries
	Year-on-year Cash Flow Growth Rate		Average Daily Turnover Rate of Investment-oriented Tertiary Industries
	Year-on-year ROE Growth Rate		PE Median of Investment-oriented Tertiary Industries
Technical Innovation Capability	Intangible Assets		Actual Trading Days in the Interval
	Proportion of Technical Employee		Average Value of Backward Adjusted Stock Price in the Interval
	Proportion of R & D Expenditure to Sales		Price-quantity(PQ) Average in the Interval
Human Resources	Total Number of Employees	Pre-interval Trading	Turnover Rate in the Interval
	Proportion of Executives	Enterprise Valuation	PB Change
	Proportion of Independent Directors		PE Change
	Proportion of Employees with Bachelor Degree or Above		Basic EPS
Corporate Governance Efficiency	Ownership Concentration		Diluted EPS
	The Sum of the Shareholding Ratio of the Top 5 Shareholders		Book Value per Share(BPS)
	The Sum of the Shareholding Ratio of the Top 10 Shareholders		PETTM
	Types of Major Shareholders		Operating Cash Flow per Share
External Influence	Practicing Quality Evaluation of Sponsored Securities Firms		
	Number of Market Makers		
	Amount of Government Subsidy		
	Total Financing		
	Amount of Placement and Allotted Shares		
	Latest Financing Time		
	Number of Historical Investors		
	Number of Tags		

3. Design and Construction of Valuation Model

3.1 Design of Valuation Model

Based on the constructed influencing factor system in Section 2, we design an ensemble learning based valuation model for enterprises listed on the new OTC market. Specifically, we take XGBoost and Random Forest (RF) as the base models, and adopt stacking ensemble as the ensemble strategy. The design process of the ensemble model is shown in Figure 1.

First, we select a subset of features from the constructed factor system by recursively eliminating

the features with the least contribution for the prediction of the target variable. Then, these features are used to construct the XGBoost prediction model and the RF prediction model. Specifically, we use GridsearchCV and 5-fold cross-validation to tune the parameters of the two models, select the optimal parameters and verify the models. Finally, we construct the XGBoost-RF ensemble model by stacking ensemble of the XGBboost model and the RF model with the optimal parameter settings.

3.2 Pre-Processing of Feature Date

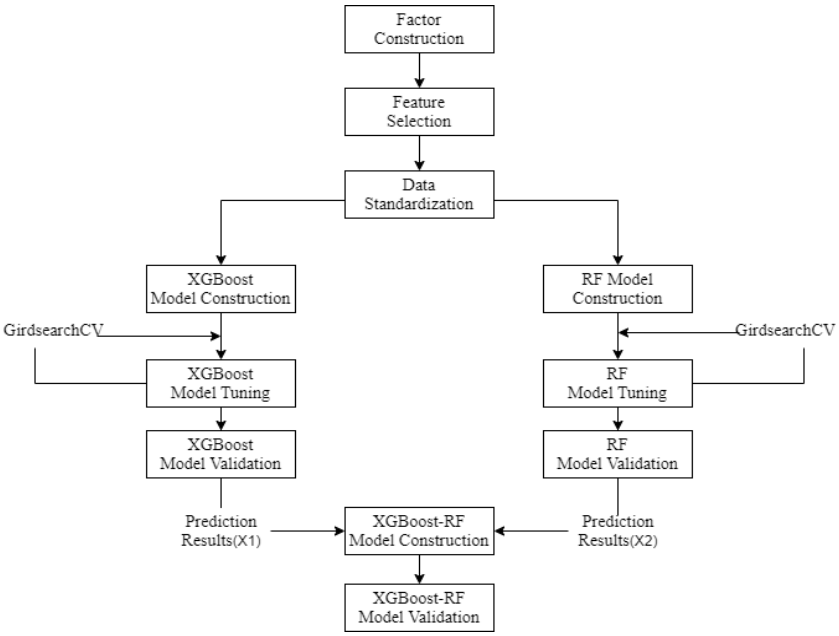

Figure 1: Design of XGBoost-RF fusion model

As the data of the factors affecting the market value of the new OTC market listed enterprises may have errors and missing values. To enhance the accuracy of the valuation, we should first deal with the missing values reasonably. For different cases, we adopt different methods to deal with the missing.

(1) Deletion: if the sample feature value is missing more than 50%, we directly delete the feature. Otherwise, we fill the missing values by some methods.

(2) Mean value filling: for numerical features, we fill the missing values with the mean value of the sample with the same feature.

(3) C4.5 method: by searching for the relationships between features, the missing features are filled in by the transformation of other features;

(4) Fixed value filling: according to the actual meaning of the feature, using the zero value or a specific feature value to fill the missing values.

(5) Previous period data filling: for time series features, using the data of the previous period to fill missing values in the current period.

(6) Construct derivative variable: define a new binary variable to indicate whether the feature value is missing.

In the research process, according to the actual situation of the collected empirical sample data set, the missing values of different features are analyzed first, and then processed in a reasonable way.

3.3 Evaluation Methodology and Metrics

To avoid over-fitting and improve the generalization ability of the model, we adopt 5-fold cross-validation to divide the sample data set. The detail process is described as Figure 2.

First, the sample data set is randomly divided into 5 equal subsets without replacement. Then, each subset is used at the test data once, and the remaining four subsets are combined as the training data. At last, the average values of the 5 cross-validation are used as the final results.

To evaluate the performance of the proposed model, we adopt two widely indicators as the evaluation metrics, i.e., the goodness of fit (R^2) and mean square error (MSE).

Let y_i and $\hat{y}_i$ be the true value and the predicted value of the i-th sample respectively. Let $\bar{y}$ denote the sample mean, then

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}, \quad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad MAE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

R^2 reflects the predictive effect of the model. The closer it is to 1, the better the prediction effect is. MSE is used to measure the error between the predicted value and the true value. The smaller the MSE, the better the model is.

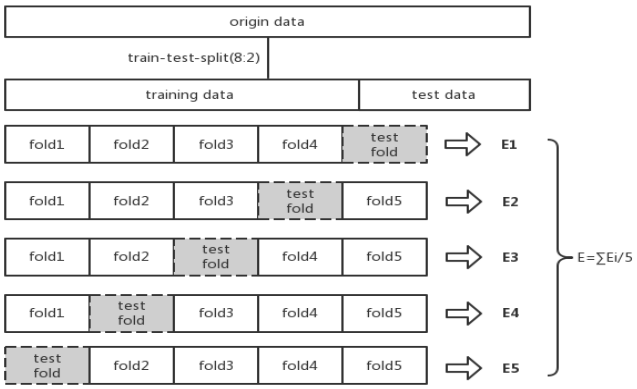

Figure 2: Flow chart of 5-fold cross-validation

3.4 Design of Target Variable

The market value of an enterprise is usually measured by the transaction price of its stock at the latest point in time. However, since the enterprises listed on the new OTC market may have no transaction records for a long time and the limitation of stock price fluctuation is relatively wide, it is not appropriate to use the stock price at the latest time point as the target variable. Therefore, we take the average value of stock transaction prices within a certain period of time as the target variable of the valuation model.

In addition, considering that the enterprise listed on the new OTC market will pay dividends to shareholders, the stock price will change after excluding right and dividend, while the actual value does not change. Therefore, to measure the stable value of enterprise listed on the new OTC market over a period of time, we take the average value of all backward adjusted transaction prices within a period of time as the target variable of the valuation model.

3.5 Feature Selection

The feature selection in this paper is divided into two stages.

First, the XGBoost algorithm is used to carry out regression analysis on feature variables and the target variable to verify the correlation between features and the target variables.

Then, according to the XGboost decision trees constructed in the first stage, the promotion of each feature of each tree in the division criteria is calculated, and then all the trees are aggregated to get the feature weights. The corresponding features are removed iteratively according to the feature weights from small to large. After the removing of each feature, the XGboost algorithm is used to calculate the goodness of fit of model with the retained features. Finally, the feature subset with the highest goodness of fit is selected, which are shown in Table 2.

Table 2: Feature Subset with the Highest Goodness of Fit

Categories	**Influencing Factors**	**Categories**	**Influencing Factors**
Profitability	Gross Profit Margin	Pre-interval Trading	Actual Trading Days in the Interval
	Net Profit Margin		
	ROE		Average Value of Backward Adjusted Stock Price in the Interval
	Return on Invested Capital		
Enterprise Scale	Total Assets		
	Total Equity		Turnover in the Interval
	Operating Revenue		Turnover Rate in the Interval
Growth Opportunity	Year-on-year Cash Flow Growth Rate	Enterprise Valuation	PB Change
			PE Change
Corporate Governance Efficiency	The Sum of the Shareholding Ratio of the Top 10 Shareholders		Basic EPS
			Diluted EPS
			Book Value per Share(BPS)
External Influence	Number of Market Makers		PETTM
Operating and Debt Paying Ability	Current Ratio		Operating Cash Flow per Share

3.6 Construction of Valuation Model

3.6.1 XGBoost model construction

With the selected feature set and the designed target variable, we first construct two base models, XGBoost model and RF model. Then, we construct the XGBoost-RF ensemble model by stacking integration with the two base models.

The objective function of the XGBoost model is defined as follows,

$$\min Obj = L(\theta) + \Omega(\theta),$$

where $L(\theta)$ denotes the loss function, which is used to measure the fitting effect of the model, and $\Omega(\theta)$ is the regularization term, which is used to measure the complexity of the model. Specially, we adopt the square loss function as the loss function $L(\theta)$, i.e.

$$L(y, f(x)) = \frac{(y - f(x))^2}{2}.$$

And, L2 regularization is used as the regularization term $\Omega(\theta)$.

3.6.2 RF model construction

Random forest is composed of multiple CART regression trees. First, for each CART regression

tree, the least square method is used for model fitting. Then, the average method is used to integrate the output results of all CART regression trees. Finally, the effects of the model is measured by the values of R^2, MSE and MAE.

3.6.3 Construction of XGBoost-RF fusion model

The RF model and the XGBoost model have different strengths, which can be combined to make better predictions. The RF model uses data parallelization to generate the CART regression trees. Therefore, it has a strong generalization ability. For the XGBoost model, there is a strong dependency between individual learners, and the final strong learner is generated through serialization, which has a strong model fitting ability. To combine the strong fitting ability of the XGBoost model and the strong generalization ability of the RF model, we construct an ensemble model with these two base models as the valuation model of the new OTC market listed enterprise. Specifically, we adopt the Stacking ensemble strategy to construct the XGBoost-RF ensemble valuation model.

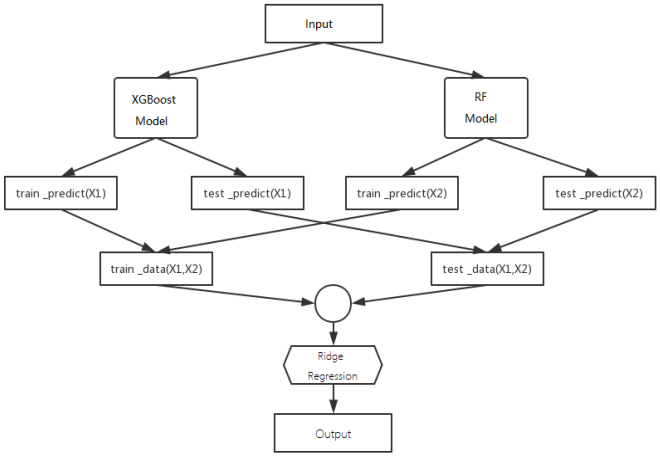

Figure 3: Stacking integration process

Stacking ensembleisa strategy to use the outputs of base models, e.g. the RF model and the XGBoost model, as features, and use another learning model to make the final predictions. In this paper, the ridge regression with strong stability and high explain ability is adopted as the second layer model.

When constructing the second layer Ridge regression model, the GBDT model's prediction result for each sample is taken as X1, the RF model's prediction result for each sample is taken as X2, and the true target variable is denoted as Y.

The stacking ensemble process is shown in Figure 3. The stacking ensemble method uses the XGBoost model and RF model's prediction results train_predict as the training set train_data of the Ridge regression model. After the Ridge regression model is trained through the train_data, its prediction value of test_data is compared with the true Y value to verify the effect of the XGBoost-RF ensemble model.

4. Empirical Analysis of Valuation Model

4.1 Sample Data Selection and Conversation

To evaluate the performance of the proposed valuation model for enterprises listed on the new OTC market, we selected the more than 2000 high-tech enterprises recommended by Wind

Information in the concept category of the new OTC market as the sample set. The target variable of the valuation model is selected as the average value of all backward adjusted transaction prices within a period of time. The feature variables is taken as the indicators of the influencing factor system constructed in Section 2.The sample interval is divided by half a year from 2016 to 2018. In order to ensure the validity of the average backward adjusted transaction prices in the interval, the stocks are required to have transaction records in the corresponding interval and have a certain trading volume. Therefore, it is necessary to screen out samples with a turnover rate less than 5%. After the screening, 1494 samples are retained, including 478 enterprises in IT industry, 407 enterprises in industry, 199 enterprises in raw material industry, 163 enterprises in non-daily consumer goods industry, 109 enterprises in health care, and 138 enterprises in the other six industries. The retained samples are in line with the distribution of investment-oriented primary industries.

The feature variables of the constructed influencing factor system include different kinds of indicators, e.g., absolute indicators, relative indicators, and percentage indicators. The magnitude orders of these feature variables are quite different. To eliminate these negative influence, we need to standardize the data before input them into models. The commonly used data conversion methods include Z-score standardization and min-max normalization. In order to eliminate the influence of noise introduced when filling missing values in the data processing process, we adopt the z-score standardization method, and minimize the influence of noise point data on the model results through the method of mean variance centralization. The z-score standardized conversion method is as follows:

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x_j}}{S_j},$$

Where $\bar{x}_j$ and S_j are the sample mean and standard deviation of the j-th feature variable, respectively.

4.2 Comparison Analysis of Model Results

In the following, we will compare the multiple linear regression model, XGBoost model, RF model and the XGBoost-RF fusion model integrated by stacking. Through the index goodness of fit R^2, mean square error MSE and mean absolute error MAE, the validity and accuracy of the model in the market value evaluation of new OTC market listed enterprises are verified. The comparison of final model results is shown in Table 3.

Table 3: Comparison of model results

Model	R^2	MSE	MAE
Multiple Linear Regression Model	74.45%	44.42	3.88
XGBoost Model	93.48%	29.17	3.13
RF Model	80.07%	54.82	3.51
XGBoost-RF Fusion Model Integrated by Stacking	94.26%	27.05	3.13

From the evaluation index of the model, we can see that the XGBoost-RF fusion model based on stacking integration mode has high accuracy and low error, which shows that it can combine the strong fitting ability of XGBoost model and the strong generalization ability of RF model, and has good performance in the empirical sample set of this paper. This model significantly improves the accuracy of the new OTC market valuation model.

5. Conclusion

To address the insufficient factor system of existing research, we first constructed a more comprehensive and effective value influencing factor system for evaluation of enterprises listed on the new OTC market. Then, to mining the non-linear relationship between features and the enterprise market value, we constructed aXGBoost-RF ensemble model based on XGBoost and RF to study the valuation of new OTC market listed enterprises. Empirical analysis on 1494 enterprises verified the rationality and accuracy of the constructed influencing factor system and the proposed valuation model for enterprises listed on the new OTC market. With the continuous development and improvement of the regulatory system of the new OTC market, we will study new influencing factor that may affect the market values of enterprises listed on the new OTC market in the future.

Acknowledgements

This work was supported by General Program of Beijing Association of Higher Education (NO.MS2023286).

References

[1] R. Aggarwal, TheFama-French Three Factor Model and the Capital Asset Pricing Model: Evidence from the Indian Stock Market [J], Indian Journal of Research in Capital Markets, 2017, 4(2): 36-47.
[2] E. I. Altman, Financial ratios discriminant analysis and prediction of corporate bankruptcy [J]. Journal of Financial Economics, 3(1976): 145-166.
[3] F. Black and M. Scholes. The Pricing of Options and Corporate Liabilities [J]. Journal of Political Economy, 1973, 81(3): 637-654.
[4] S. Boubaker, H. Mansali and H. Rjiba, Large controlling shareholders and stock price synchronicity [J]. Journal of Banking and Finance, 2014, 40: 80-96.
[5] Y. Chen, Research on Enterprise Valuation in Venture Capital [J]. Financial Theory & Practice, 2010(01): 64-67.
[6] S. W. Choi, An Empirical Study of Capital Asset Pricing Model and Fama-French Three-Factor Model [D]. UCLA, 2017.
[7] V. T. Datar, N. Y. Naik and R. Radcliffe, Liquidity and stock returns: An alternative test [J]. Journal of Financial Markets, 1998, 1(2): 203-219.
[8] J. L. Davis, E. F. Fama and K. R. French. Characteristics, Covariances, and Average Returns: 1929 to 1997 [J]. The Journal of Finance, 2000, 55(1): 389-406.
[9] R. Fang, Research on valuation theory of Listed Enterprises on China's New OTC Market [J]. Times Finance, 2017(26): 181-182.
[10] X. Guo and L. Xiong, Factor empirical test of multi-dimensional financial indicators of listed enterprises on the new OTC Market, Financial Supervision [J]. 2012(32): 28-32.
[11] J. Lin, M. Wang and L.Cai, Are the Fama–French factors good proxies for latent risk factors? Evidence from the data of SHSE in China [J]. Economics Letters, 2012, 116(2): 265-268.
[12] F. Modigliani and M. H.Miller. The Cost of Capital Corporation Finance and the Theory of Investment [J]. The American Economic Review, 1958, 48(3): 261-297.
[13] W. Qin, A Study on the Valuation Method of Chinese NEEQ [D]. Huazhong University of Science and Technology, 2017.
[14] W. F. Sharpe, Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk [J]. The Journal of Finance, 1964, 19(3): 425-442.
[15] Y. Wang, The Research of the Method about the New Three Board Corporate Valuation [D]. Tianjing University, 2016.
[16] Q. Yan and J. Tao, Research on Performance Evaluation of Listed Enterprises on the New OTC Market, Finance and Accounting Monthly, 2014(04): 13-16.
[17] Y. Zhang, Research on Valuation of New OTC Market Listed Company—Take the Internet Industry for Example [D], Zhejiang University, 2017.