

Application of Factor Sensitivity Analysis Using Random Forest Model in Investment Decisions

– Taking the Semiconductor Sector for Example

Wei Jiang^{1,a,*}

¹*Institute of Economics, Zhejiang Wanli University, Ningbo, China*

a. 2227443162@qq.com

**corresponding author*

Abstract: The semiconductor industry is the key field of the country's scientific and technological innovation and industrial transformation and upgrading. Sensitivity analysis by random forest model can reflect the influence of input variables on the output response. Through sensitivity analysis, 32 financial factors will be considered and the weight coefficients affecting investors' investment decisions in the semiconductor industry will be determined and ranked using the random forest model. The research results show that mean deviation, Research & Development expense and other factors play a high weight in the investment decision of semiconductor companies. Based on the above research, it is effectively judge the profitability of all listed companies in the semiconductor industry according to the factors in the secondary factor pool. The application of the Random Forest model in the semiconductor sector provides valuable insights into key factors influencing investment returns.

Keywords: machine learning, random forest, investment decisions, sensitivity analysis.

1. Introduction

As the top product of the science and technology industry, the semiconductor industry is the key field of the country's scientific and technological innovation and industrial transformation and upgrading. In recent years, with the policy support, financial encouragement and the impact of geopolitics, the semiconductor industry is clearly thriving. For example, in 2020, the State Council have put forward that we will continue to implement value-added tax preferential policies and corporate income tax policies for integrated circuit enterprises and software enterprises, including design, equipment, materials, packaging, testing enterprises and software enterprises[1]. However, In terms of industry stock selection, it is usually difficult for retail investors to screen and extract truly effective data from various financial indicators to help them make investment decisions and make their expected revenue. Moreover, because each financial data points to different company capabilities, the degree of correlation to profitability is also different.

Sensitivity analysis can reflect the influence of input variables on the output response. The sensitivity analysis includes local sensitivity and global sensitivity analysis.[2]The index use sensitivity analysis is an important parameter calculated from the stock data to predict the trend of

the stock market. It is also a tool often used by investors.[3]The local sensitivity can respond to the influence of input variables on the characteristics of output at the nominal value. The global sensitivity analysis, known as the importance measure analysis, can estimate the influence of input variables in the whole distribution region on the characteristics of output. Sensitivity analysis can help customers identify key variables. In terms of investment decisions, sensitivity analysis can determine which financial data factors have the greatest impact on a company's earnings, so as to help decision makers focus on these factors for further research and optimization of their investment decisions.

However, a single machine learning algorithm may have some shortcomings under sensitivity analysis. First, the fitting ability of a single machine learning algorithm is weak, Especially in the face of noise, outliers or missing values, the output of the model may fluctuate greatly, thus affecting its stability and reliability. Furthermore, when faced with high-dimensional data or multicollinearity problems, a single algorithm may be difficult to effectively capture the true pattern of the data, resulting in increased uncertainty of the results. When dealing with complex and non-linear relationships, the performance of a single algorithm may not be satisfactory, and it is susceptible to changes in data characteristics.

Random forest (RF) is composed by multiple decision trees (DTs), it is an ensemble learning method proposed by Breiman[4].RF are considered one of the best out-of-the-box classification and regression algorithms due to their high level of predictive performance with relatively little tuning. Pairwise proximities can be computed from a trained random forest and measure the similarity between data points relative to the supervised task[5]. RF has many advantages, such as strong robustness, good tolerance to outliers and noise. RF has a wide range of application prospects, such as geographical energy, chemical industry, health insurance and data science competitions. RF cannot only deal with classification and regression problems but also analyze the critical measure. RF provides two kinds of importance measures: Mean Decrease Impurity (MDI) based on the GINI index and Mean Decrease Accuracy (MDA) based on Out-of-Bag (OOB) data[6].

As ensemble learning, the random forest model can effectively process high-dimensional data, such as constructing each decision tree by randomly selecting a subset of features, thus reducing the computational complexity and correlation between features. Besides, because random forest models make final decisions by averaging the predictions of multiple decision trees, the risk of overfitting a single model is reduced. This integrated approach gives the model a better ability to generalize when dealing with new data.

The parameters in a Random Forest model each play a crucial role. The training ratio determines the proportion of data used for training, with 0.7 indicating that 70% of the data is for training and the remaining 30% is for validation or testing. Data shuffling ensures that the dataset is randomly ordered before training, preventing any sequence bias and improving the model's generalization ability. The node splitting evaluation criterion, such as the GINI index, is used to select the optimal split by assessing node impurity, thereby reducing classification errors. The number of decision trees refers to how many trees are included in the model, with more trees enhancing stability and accuracy. Sampling with replacement allows data to be repeatedly sampled during tree training, increasing diversity and boosting robustness. The minimum sample number for internal node splitting sets the minimum sample requirement for a node to be eligible for splitting, helping to prevent overfitting. The minimum sample number for leaf nodes specifies the least number of samples a leaf node should contain, ensuring model flexibility. Maximum tree depth limits the number of layers in a decision tree, controlling complexity and preventing overfitting. Lastly, the maximum number of leaf nodes restricts the number of leaf nodes in each decision tree, further simplifying the model to improve its performance on new data.

2. Materials and methods

2.1. First-order factor pool—three dimensions

Based on the JoinQuant database, volume and price factors, risk model factors, and financial fundamentals are chosen from, subdividing the dimensions, and eliminating the relevant indicators with incomplete data and multicollinearity. Finally, 32 related indexes were selected to construct the first level factor pool, and the second level was extracted by random forest model algorithm factor pool.

2.1.1. Volume-price factor

Considering the availability and timeliness of stock data, we are choosing fundamental factors and risk model factors

Based on the addition of volume and price factors, it can capture the latest market dynamics in a more timely manner, and eliminate macro

The impact of events and alternative factors on the market. The volume price factor is based on the technical level of stock analysis and extension

The essence of factor model is to describe the fixed behavior pattern of people based on behavioral finance that can improve the trend of stocks

Line judgment and prediction. The volume and price factors decompose the dimensions of emotion, risk, technology and momentum.

The emotional factors include the trading volume, the moving average of trading volume, the dispersion of William's variation and the standard of trading amount.

The risk factors include Sharpe ratio, annualized return variance, and individual stock return skewness.

Technical factors include money flow indicators, upper/lower track (Bollinger line) indicators, exponential moving averages(EMA), and Moving Average.

Moving average;

Momentum factors include mean difference, divergence, cumulative energy line and momentum line.

2.1.2. Risk model factors

It can describe the rotation characteristics of the industry and estimate the risk of the equity portfolio, and the risk model factors on the style and industry dimension

Because the research direction of this paper is limited to semiconductor industry, the style factor is selected for analysis.

Style factors include market cap, liquidity (turnover rate), beta, residual volatility, leverage, non-linear market cap, book-to-market ratio.

2.1.3. Financial fundamental factors

It is able to describe the financial status of a company and measure its operating status and growth potential based on important indicators from the profit and loss statement, cash flow statement, and balance sheet. The financial fundamentals factor has decomposed the quality, foundation, growth, and earnings per share dimensions based on key indicators.

The quality factor includes net profit margin, owner's equity ratio, shareholder's equity ratio, non-current asset ratio, net profit attributable to parent company shareholders, and net profit after excluding non-recurring losses; the foundation factor includes net operating capital, revenue, research

and development expenses, selling expenses, net profit, operating costs, financial expenses, and net debt; the growth factor includes revenue growth rate, return on equity ROE, return on assets ROA, price-earnings ratio, price-to-book ratio, price-to-sales ratio, and price-to-cash-flow ratio; the earnings per share factor includes earnings per share EPS, net asset per share, operating profit per share, and operating revenue per share TTM.

Table 1: First-order factor pool

number	factor	number	factor
X1	Market Value	X2	Turnover Rate
X3	P/E	X4	P/B Ratio
X5	Price-to-Sales Ratio	X6	P/CF Ratio
X7	Net Working Capital	X8	Net Debt
X9	D/E Ratio	X10	Non-Current Assets Ratio
X11	Equity Ratio	X12	Operating Revenue
X13	Operating Costs	X14	Sales Expenses
X15	Administration Costs	X16	Financial Expenses
X17	Operating Profit	X18	Net Profit
X19	Net Profit Attributable to Shareholders of the Parent Company	X20	EPS
X21	Net Profit Excluding Non-Recurring Gains and Losses	X22	ROE
X23	ROA	X24	Net Profit Margin
X25	Revenue Growth Rate	X26	Momentum Line
X27	R&D Expense	X28	Accumulated Energy Line
X29	Mean Deviation	X30	EMA
X31	Moving Average	X32	Percentage Deviation

2.2. Research Methods

Based on GINI coefficient, this paper calculates the features of downward splitting of each CART decision tree, selects the features with the highest GINI coefficient to be incorporated into the training set data, constructs a random forest model with 100 CART decision trees, and implements the algorithm with Python. Each characteristic value is output, that is, the importance of the factors in the first-level factor pool is sorted and extracted into the second-level factor pool.

2.2.1. Build data set

Factor data of all stocks in the semiconductor industry from 2021-2023 were crawled as sample data, and eigenvalue selection of random forest was carried out. The first-level factor pool predicts the selected stocks. To construct the data set, the author marks individual stocks as 1 if the predicted income level exceeds the average level, and as 0 otherwise.

a. data pre-processing

The data of individual stocks are processed centrally, some missing and wrong data are eliminated and corrected, and some factor data with weak stability are treated with Z-SCORE standardization.

b. Random forest fitting

Formula for calculating the GINI coefficient $GINI(P) = 1 - \sum (P_i^2)$, where P_i represents the proportion of class i samples to calculate the GINI coefficient of each feature, so as to measure the

purity of the data set and reduce the impurity of the fission node to the minimum value. The permutation importance value is obtained, and the eigenvalues with high GINI coefficient are selected as the input of the model, and the importance of the eigenvalues is sorted according to the GINI coefficient from large to small. This group has chosen the following key model parameters: a training ratio of 0.7, data shuffle, a GINI node splitting evaluation criterion, a decision tree number of 100, reverse sampling, a minimum sample number of internal node splitting of 2, a minimum sample number of leaf nodes of 1, a maximum tree depth of 10, and a maximum number of leaf nodes of 50.

Table 2: Model label type to proportion

Data Set	Variety	Count	Percentage
Training Set	Class_0	50	0.7042
	Class_1	21	0.2958
	Sum	71	100%
Testing Set	CLASS_0	23	0.697
	CLASS_1	10	0.303
	Sum	33	100%

Table 3: Model prediction result

predicting results	CLASS	Revenue growth rate	Mean Deviation	Turnover Rate	Predicted outcome probability_0	Predicted outcome probability_1
0	0	0.93	-0.48	0.62	1	0
1	0	1.65	-0.18	0.79	1	0
0	1	0.90	1.26	0.82	0.49	0.51
0	0	1.43	-1.41	0.57	1	0
1	1	0.86	-2.91	0.38	1	0
...	...	...	...	...	...	...
0	0	1.49	-0.52	0.54	1	0
1	0	0.80	-0.96	0.42	0.98	0.02
1	0	0.99	-1.67	0.89	1	0
0	1	1.22	-0.92	0.74	0.77	0.23
0	0	0.94	-1.13	0.43	1	0

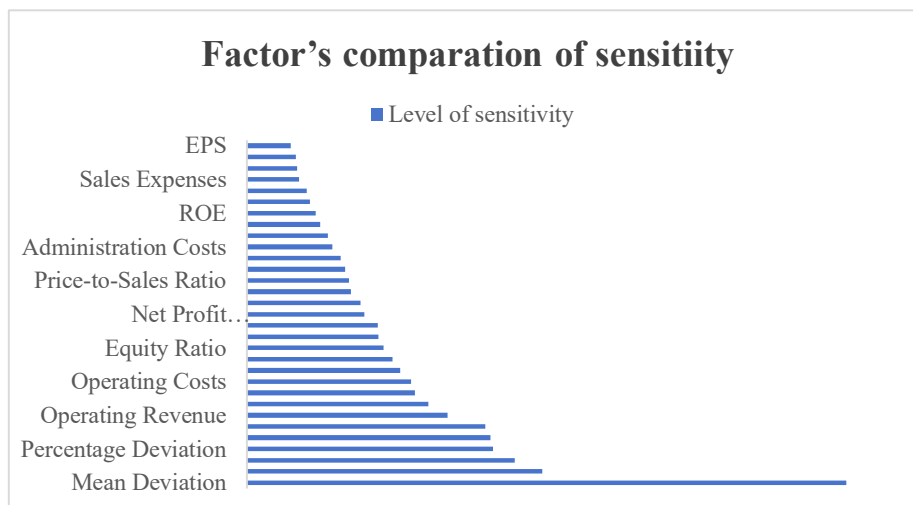


Figure 1: Factor feature value importance ranking

Thus, the weights of each factor were obtained and sorted, and the ten factors with the largest weight were selected into the secondary factor pool.

Table 4: Operating Costs

Number	Factor	Weight
X1	Mean Deviation	0.10911
X2	R&D Expense	0.053722
X3	momentum line	0.048812
X4	Percentage Deviation	0.044766
X5	Revenue Growth Rate	0.044286
X6	Non-Current Assets Ratio	0.043402
X7	Operating Revenue	0.036488
X8	Turnover Rate	0.033067
X9	D/E Ratio	0.030624
X10	Operating Costs	0.029933

3. Results and Discussion

Based on the data in Table 4 derived from the random forest model, we can very significantly understand that the ten factors that account for the largest weighting coefficients of investment decisions in the semiconductor industry are respectively mean deviation, R&D expense, momentum line, percentage deviation, Revenue growth rate, non-current assets ratio, operating revenue, turnover rate, D/E ratio, operating costs. What's more, the weight coefficient of the mean deviation reached 0.10911, which far exceeded the other factors.

Businesses use mean deviation as a statistical tool to analyze the volatility, trend, and consistency of data in a corporate setting. It plays an important role in decision-making, financial management, market analysis, quality control and many other fields.

Based on the above research, it is effectively judged the profitability of all listed companies in the semiconductor industry according to the factors in the secondary factor pool. According to the feature importance and sensitivity analysis, several investment strategies can be formulated. Firstly, prioritize investments in companies with favorable mean deviation and high R&D expenses. Furthermore, companies with strong revenue growth trends are likely to offer better returns. By incorporating these insights, investors can make more informed decisions and optimize their portfolios in the semiconductor sector.

However, the model's training revealed lower-than-expected accuracy on the test set, leading to overfitting. The author suspected errors in data cleaning due to the stock's sensitivity to unit changes and large number of outliers.

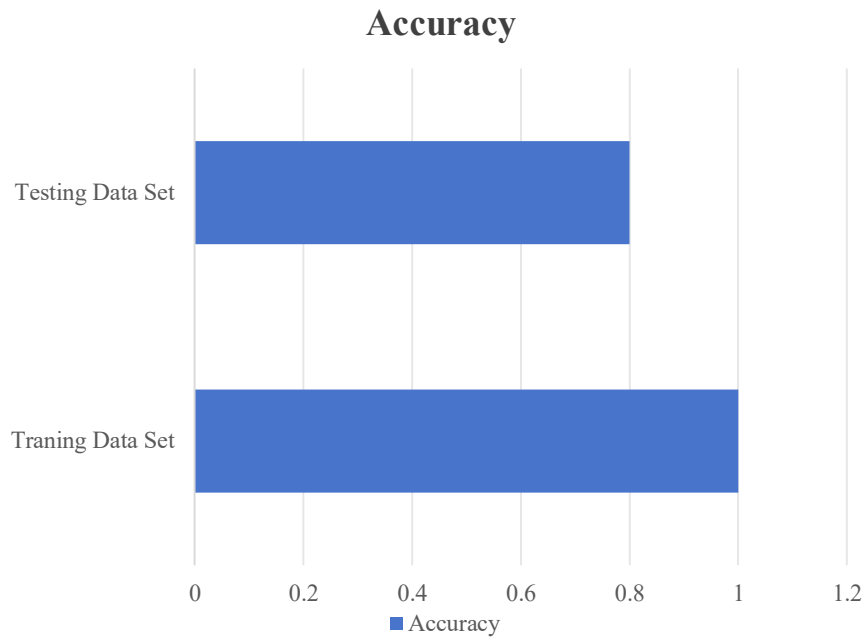


Figure 2: Data Set Accuracy

The application of Z-scores has certain limitations, especially when the data distribution does not conform to a normal distribution. Firstly, Z-scores assume that the data is approximately normally distributed. However, many datasets do not meet this assumption in practice. If the data distribution deviates significantly from normality, the Z-score may not accurately reflect the true deviation of data points. Secondly, Z-scores are not applicable to categorical or discrete data, as these types of data do not have a meaningful mean or standard deviation, making the calculation of Z-scores irrelevant. Additionally, when the data sample is small, estimates of the mean and standard deviation may be unreliable, leading to inaccurate Z-score calculations. In cases where there is insufficient data, Z-scores may fail to effectively standardize the data, reducing their usefulness in analysis.

As a result, missing values can be handled through forward filling, backward filling, or interpolation methods, or by deleting rows with few missing values if necessary. Secondly, using statistical methods like Z-score or box plots to identify and address outliers ensures data accuracy. Standardization or normalization helps eliminate scale differences in the data, making model training more effective. Additionally, data smoothing techniques such as moving averages can reduce noise and provide clearer price trends. Interpolation methods can fill gaps for non-trading periods, maintaining data continuity. Consistency checks ensure that timestamps and data frequencies are uniform, avoiding analysis bias. Finally, identifying and separating the impact of major market events helps prevent them from skewing analysis results.

4. Conclusion

The findings from this study demonstrate the potential of using Random Forest models for factor sensitivity analysis in investment decision-making, particularly in the semiconductor sector. However, there are several avenues for future research that could further enhance the model's applicability and robustness:

Firstly, while this study focused on the semiconductor sector, applying the Random Forest model to other sectors such as technology, healthcare, or energy could provide valuable cross-sector insights.

This would also allow for comparative analysis across different industries, helping investors diversify their portfolios more effectively.

Secondly, the current model primarily relies on traditional financial and macroeconomic data. Future research could explore the integration of alternative data sources, such as social media sentiment, news analytics, or even satellite imagery, to capture additional market signals. These alternative datasets could further enhance the model's predictive power and provide early indicators of market shifts.

Third, although Random Forest is a robust and powerful model, combining it with other machine learning algorithms in an ensemble learning framework could potentially improve accuracy and stability like Gradient Boosting Machines (GBMs), Support Vector Machines (SVMs), or deep learning approaches could lead to better performance.

Finally, future research could explore how the insights from the Random Forest model can be directly integrated into portfolio optimization algorithms. This could involve the development of new optimization techniques that leverage the factor importance and sensitivity results to construct more efficient and diversified portfolios.

By addressing these areas, future work can build on the foundation established in this study, enhancing the utility of machine learning models in investment decision-making and further bridging the gap between advanced analytics and practical financial applications.

The application of the Random Forest model in the semiconductor sector provides valuable insights into key factors influencing investment returns. The model's robustness and interpretability make it a powerful tool for enhancing investment strategies. By leveraging the findings from factor sensitivity analysis, investors can make more informed and strategic decisions.

References

- [1] State Council (2020.8). *Several policies for Promoting high-quality Development of Integrated Circuit Industry and Software Industry in the New Era*(in Chinese)
- [2] Lu, Z. Z., Li, L. Y., Song, S. F., Hao, W. R. (2015). *Theory and solution of importance analysis for uncertain structural systems*, pp. 1–5. Beijing: Science Press (in Chinese)
- [3] Lili Y, Benling L, Peng L, Rubo Z. (2023). *Research on stock trend prediction method based on optimized random forest*
- [4] reiman, L. (2001). *Random forest*. *Machine Learning*, 45(1), 5–32. DOI <http://dx.doi.org/10.1080/14498596.2017.1367331>
- [5] Loecher, M. (2020). *From unbiased MDI feature importance to explainable AI for trees*. <https://www.researchgate.net/publication/340224035>
- [6] Rhodes, JS., Cutler, A., Moon, KR.(2023.). *Geometry- and Accuracy-Preserving Random Forest Proximities*