Enterprise Investment Value Analysis Based on Machine Learning Model of Rapidminer

Zhang Chong*, Zhang Xinrui and Yang Zipei
College of Information Science, Beijing Language and Culture University, Beijing, China

*Email: zhangchong@blcu.edu.cn

Abstract. Whether an enterprise has investment value and how to measure the investment value of an enterprise are the primary concerns of every investor. Effective evaluation of enterprises’ investment values can not only help enterprises to recognize their own problems and values and to achieve faster growth, but also help investors to obtain more investment returns. The investment value of enterprises is reflected both in financial data and non-financial data. This paper is based on data mining and analysis of 3500 listed companies to establish a more comprehensive and objective evaluation model of enterprise investment value. Also, after data cleaning, data are trained by machine learning in RapidMiner software to compare the accuracy of several models, thus key factors affecting enterprise investment value are analyzed and investment advice will be given to investors, and enterprises can strengthen their capabilities in concerning areas to enhance their own value to attract more investment.

1. Introduction
The growing stock market calls for investors to apply more scientific analysis methods to identify the investment value of enterprises and the best investment solution. This paper studies the investment value of enterprises based on the operating data of 3500 listed companies. Financial and non-financial indicators are included to establish a more scientific and comprehensive evaluation system. Also, RapidMiner is used to run multiple machine learning models on the standardized data and compare the accuracy of six models on the test set to analyze the impact indicators of enterprises’ investment value to provide investors with a basis for decision-making.

2. Construction of evaluation system of enterprise investment value indicators
Based on Benjamin Graham's investment analysis philosophy [1] about the investment value of enterprises, this paper adopts the traditional definition that the assessed enterprise has estimated value for investors with investment targets.

Foreign scholars have done deep research on the investment value of enterprises by making financial analysis and formed classic assessment models and methods, such as Discounted Cash Flow Model, Economic Value Added Model, Dividend Discount Model, Finance Index Analysis, etc.

After obtaining relevant enterprise data, domestic and foreign scholars have adopted statistical or machine learning methods like Principal Component Analysis, Factor Analysis, Correlation Coefficients and Support Vector Machine algorithms to establish evaluation systems, leading to high accuracy of the prediction of affecting indicators and classifications [2].
After comparing the predecessors’ selections of financial indicators and their results, this paper combines relevant theories to design an evaluation system based on the following principles:

- **Scientific principle.** Each indicator needs to measure the capability of a certain aspect of the enterprise, and the coverage of indicators should not overlap [3].
- **Objectivity principle.** The selection should try to find more objective data in the original data, which means indicators with a large number of missing values should not be included.
- **Systematic principle.** Both static and dynamic indicators are included in the research scope, and historical and future indicators are placed in the evaluation system reference together.

### 2.1. Primary selection of indicators in evaluation system

It is certain that the investment value of enterprises is affected by both financial and non-financial indicators. Financial indicators only reflect previous performance, but non-financial indicators are often future-oriented and require years of efforts to achieve improvement. Once successfully fulfilled, they will significantly improve the enterprise’s performance. Therefore, these non-financial indicators can promote managers to focus on improving the overall performance of the enterprise.

The indicators included in the test data are generally divided into two aspects, namely financial indicators and non-financial indicators, and several relating secondary indicators are selected. The initial construction of indicators in evaluation system is shown in Table 1.

| Indicator type        | Secondary indicators                                      | Selected financial indicators                                      |
|-----------------------|----------------------------------------------------------|-------------------------------------------------------------------|
| **Financial indicators** |                                                          |                                                                    |
| Profitability        | Return on Equity                                         | Diluted Return on Total Assets                                     |
|                       |                                                           | Gross Margin                                                      |
|                       |                                                           | Net Margin                                                       |
|                       |                                                           | Basic Earnings Per Share(Yuan)                                    |
|                       |                                                           | Net Asset Value Per Share(Yuan)                                   |
|                       |                                                           | Reserve Fund Per Share(Yuan)                                      |
|                       |                                                           | Undistributed Profit Per Share (Yuan)                              |
| Operation             | Total Assets Turnover                                    | Days Sales Outstanding                                            |
|                       |                                                           | Days Sales of Inventory                                            |
| Solvency (Financial risk indicator) |                                                        | Debt Asset Ratio                                                  |
|                       |                                                           | Current Debt Ratio                                                |
|                       |                                                           | Current Ratio                                                     |
|                       |                                                           | Quick Ratio                                                      |
| Growth                | Operating Income Growth Rate                             | Attributable Net Profit                                           |
|                       |                                                           | Rolling Revenue Growth                                            |
| **Non-financial indicators** |                                                      |                                                                    |
|                       | Number of Information Products                           |                                                                    |
|                       | Type of Enterprise                                      |                                                                    |
|                       | Business Status                                         |                                                                    |
|                       | Age of Enterprise                                       |                                                                    |
|                       | Number of Enterprise                                    |                                                                    |
|                       | Number of Patents                                       |                                                                    |

### 3. Model training and analysis based on RapidMiner

#### 3.1. Data integration and feature processing

First, data are standardized and separate tables are integrated into one normalized table. After importing data into SQL server to calculate the average value of the same indicator dimension of each
enterprise as the attribute value, the variables with a large number of missing values are removed and data with missing horizontal dimensions are completed. The uncategorized text values are numerically encoded. SPSS is used to standardize the data of each indicator except enterprise score. Then, the missing and abnormal data are processed. Import data into RapidMiner, then use X-means operator to cluster the existing data so as to fill in missing values based on similar data and Outliers operator to identify and delete the abnormal data. A total of 2904 pieces of data are left.

3.2. Feature engineering: feature selection
This paper chooses the correlation coefficient method. We first perform correlation analysis between indicators. The correlation coefficient between the “basic earnings per share” and “undistributed profit per share” indicator reaches 0.816, and the correlation coefficient between the “quick ratio” and “current ratio” indicator reaches 0.978. In order to achieve the effect of dimensionality reduction, one of the relatively relevant indicators may be retained. Therefore, after the first round of model training, the lower weighted of the two sets of correlation indicators will be deleted to optimize the model.

3.3. Introduction of RapidMiner
This paper uses six algorithms in RapidMiner Studio version 9.0 for the following data analysis. In the RapidMiner, precision and recall rates are used to measure the pros and cons of algorithms for classification problems, while root mean squared error and mean absolute error are used for regression problems. In the previous data integration, we have standardized each indicator, but still retain the original data of the enterprise value score of which the full score is 100.

3.4. Model training and analysis
First, use the split data operator in RapidMiner to split data, 60% of the data as the training set and 40% as the test set. Second, set the role of each attribute. Finally, apply model operator is used to fit the data with six algorithms: support vector machine, generalized linear model, decision tree, random forest, gradient boosting decision tree and deep learning.

3.4.1. Support vector machine. It is a kind of supervised binary classification model. It divides data into two categories by finding the optimal separating hyper plane which solves the classification problem. But in RapidMiner, it can deal with the prediction problem of continuous variables. It uses RBF as the kernel function in this paper. The performance of the model is shown in Table 2. The RMSE on the test set is 4.588. When γ is 0.01 and C is 10, the model performs optimally.

| γ    | C     | Root Mean Squared Error |
|------|-------|-------------------------|
| 10   | 10    | 5.176783                |
| 10   | 100   | 5.17652                 |
| 10   | 1000  | 5.17652                 |
| 1    | 10    | 4.961612                |
| 1    | 100   | 5.013042                |
| 1    | 1000  | 5.013042                |
| 0.1  | 10    | 4.937188                |
| 0.1  | 100   | 7.838719                |
| 0.1  | 1000  | 7.476886                |
| 0.01 | 10    | 4.588017                |
| 0.01 | 100   | 7.417322                |
| 0.01 | 1000  | 5.37213                 |
| 0.001| 10    | 4.886123                |
3.4.2. Generalized linear model. We assume that the feature attribute $x$, parameter $\theta$, conditional probability $P(y|x; \theta)$ of the dependent variable $y$ obeys $y|x; \theta \sim \text{ExpFamily}(\eta)$. The coefficient of each attribute is shown in Table 3. The RMSE on the test set is 4.715.

Table 3. Coefficients of attributes of generalized linear model.

| Attribute X               | Coefficient $\alpha$ |
|---------------------------|-----------------------|
| Number of patents $X_1$   | -0.12                 |
| Age of enterprise $X_2$   | -0.766                |
| Number of information products $X_3$ | 0.275             |
| Return on equity $X_4$    | -0.112                |
| Basic earnings per share $X_5$ | -0.075            |
| Days sales of inventory $X_6$ | -0.102            |
| Days sales outstanding $X_7$ | -0.104            |
| Attributable net profit $X_8$ | -0.077            |
| Total assets turnover $X_9$ | -0.258            |
| Reserve fund per share $X_{10}$ | -102.968        |
| Net asset value per share $X_{11}$ | -0.037          |
| Undistributed profit per share $X_{12}$ | -0.216          |
| Gross margin $X_{13}$     | 0.489                 |
| Current ratio $X_{14}$    | 0                     |
| Current debt ratio $X_{15}$ | 0.587             |
| Operating income growth rate $X_{16}$ | -0.042          |
| Rolling revenue growth $X_{17}$ | -0.181           |
| Debt asset ratio $X_{18}$ | -0.819                |
| Quick ratio $X_{19}$      | 0.185                 |
| Constant coefficient     | 81.873                |

Thus, we can conclude the function between enterprises’ investment value score $y$ and attributes:

$$y = -0.12X_1 - 0.766X_2 + 0.275X_3 - 0.112X_4 - 0.075X_5 - 0.102X_6 - 0.104X_7 - 0.077X_8 - 0.258X_9$$

$$- 102.968X_{10} - 0.037X_{11} - 0.216X_{12} + 0.489X_{13} + 0.587X_{15} - 0.042X_{16} - 0.181X_{17} - 0.819X_{18} + 0.185X_{19} + 81.873$$  \(1\)

Apparently, what have positive effects on the enterprises’ investment value score are the number of information products, gross margin, current debt ratio and quick ratio, while age of enterprises, reserve fund per share and debt asset ratio have negative effects. This result is far from our expectation, so we need to try other algorithms.

3.4.3. Decision tree. When selecting CART algorithm to deploy the model, all predicted values are the same, which means this model has poor performance, so the followed algorithms namely random forest and gradient boosting decision tree are selected to optimize the decision tree algorithm.

3.4.4. Random forest. It is an algorithm for predicting by constructing multiple decision trees. For the regression problem in this paper, the result of random forest is the mean prediction results of multiple decision trees. It uses bagging algorithm in ensemble learning to build multiple CART decision trees, and these trees form the forest [4]. The CART algorithm is still selected to deploy decision tree.

The iteration process is shown in Table 4. The smallest RMSE is 4.353, which is better than the previous three algorithms. When the tree is 60 and the depth is 7, the model performs optimally.
### Table 4. Iterative process of random forest.

| Number of trees | Maximum depth | Root Mean Squared Error |
|-----------------|---------------|-------------------------|
| 20              | 2             | 4.90788                 |
| 60              | 2             | 4.905345                |
| 100             | 2             | 4.904675                |
| 140             | 2             | 4.904877                |
| 20              | 4             | 4.626895                |
| 60              | 4             | 4.600378                |
| 100             | 4             | 4.598925                |
| 140             | 4             | 4.60446                 |
| 20              | 7             | 4.450066                |
| 60              | 7             | 4.353521                |
| 100             | 7             | 4.415709                |
| 140             | 7             | 4.409932                |

3.4.5. **Gradient boosting decision tree.** It also solves the problem by iterating multiple decision trees. It uses the negative gradient of loss function as the approximation of residuals of boosting tree algorithm to fit the regression tree [5]. The iterative process is shown in Table 5, one tree during the process is shown in Figure 1. When the number of trees is 60 and the maximum depth is 4, the model performs optimally, and the RMSE is 4.35, which has little difference from that of random forest.

### Table 5. Iterative process of gradient boosting decision tree.

| Number of trees | Maximum depth | Root Mean Squared Error |
|-----------------|---------------|-------------------------|
| 20              | 2             | 4.679394                |
| 60              | 2             | 4.516245                |
| 100             | 2             | 4.483853                |
| 20              | 2             | 4.506898                |
| 60              | 2             | 4.35278                 |
| 140             | 2             | 4.473533                |
| 100             | 4             | 4.454143                |
| 20              | 7             | 4.490369                |
| 60              | 4             | 4.49142                 |
| 140             | 7             | 4.488624                |
| 100             | 7             | 4.536015                |
| 140             | 7             | 4.556301                |

![Figure 1. One tree in the iterative process of gradient boosting decision tree.](image-url)
3.4.6. **Deep learning.** Typical activation functions in deep learning are sigmoid, softmax and Relu. Relu activation function is used in RapidMiner, and the number of the Hidden layer nodes is 50. The process on the test set is shown in Figure 2.

![Deep Learning - Predictions Chart](image)

**Figure 2.** Fitting of deep learning on the test set.

3.4.7. **Summary of algorithm results.** The performance of each algorithm is summarized in Table 6. The two algorithms with higher accuracy of the results are gradient boosting decision tree and random forest, with RMSE of 4.35 and 4.353 and MAE of 3.545 and 3.267 respectively. We can further measure the accuracy (also shown in Table 6) of these models by using the formula:

\[
\text{Accuracy} = 1 - \frac{\text{absolute value} \left( \text{actual value - predicted value} \right)}{\text{predicted value}}
\]  

(2)

| Algorithm                  | Root Mean Squared Error | Mean Absolute Error | Operating time | Accuracy  |
|----------------------------|-------------------------|---------------------|----------------|-----------|
| Gradient boosting decision tree | 4.350                  | 3.545               | 40 s           | 95.78%    |
| Random forest              | 4.353                  | 3.567               | 2 min          | 95.76%    |
| Support vector machine     | 4.588                  | 3.599               | 4 min 31 s     | 95.26%    |
| Deep learning              | 4.686                  | 3.870               | 13 s           | 95.16%    |
| Generalized linear model   | 4.715                  | 3.882               | 3 s            | 95.12%    |
| Decision tree              | 4.963                  | 4.122               | 5 s            | 95.03%    |

| Primary indicators | Index Weight | Secondary indicators            | Index Weight |
|-------------------|--------------|--------------------------------|--------------|
| Financial indicators | 4.193        | Profitability                  | 1.158        |
|                    |              | Operation                       | 0.152        |
|                    |              | Solvency                        | 2.66         |
|                    |              | Growth                          | 0.223        |
| Non-financial indicators | 0.924    | Age of Enterprise               | 0.718        |
|                    |              | Number of Patents               | 0.1          |
|                    |              | Number of Information Products  | 0.084        |
|                    |              | Registered Capital              | 0.022        |
It can be found that the weight of financial indicators is far greater than that of non-financial indicators, which means the impact of financial indicator is greater. The largest weight in financial indicators is solvency; the largest weight in non-financial indicators is the age of enterprise. Overall, the four most important indicators are: the number of patents, gross margin, current debt ratio and quick ratio. These indicators have a positive correlation with the score of enterprises’ investment value.

3.5. Model optimization

The gradient boosting decision tree algorithm is optimized below. The indicator “diluted return on total assets” with a weight of 0 is deleted. From the previous research, we have learned that the correlation coefficient between the “basic earnings per share” and the “undistributed profit per share” indicator is lower (0.816), so the two indicators are deleted. The Forward Selection operator is added and it is intended to use the forward selection method to filter features to optimize the model. Using 80% of the data as the training set and the remaining 20% as the test set, the deployed process is executed again. The RMSE reaches 4.334 now, which is lower than the previous optimal result (4.35) by 0.016.

4. Concluding remarks

Based on the analysis above, this paper draws the following conclusions: In the evaluation system of enterprise investment value indicators, the financial indicators make the most contributions and the solvency ranks first with profitability closely followed, which illustrates that liquidity of enterprise assets is the key factor in investment value. Non-financial indicators cannot be ignored though they have small weights, and of them the two most important indicators are the age of enterprise and the number of patents. Overall, the solvency of an enterprise is the key to the sustainable development of an enterprise. Profitability is the ultimate goal of an enterprise and a centralized reflection of the company's income which investors should lay great emphasis on.

This research also has certain limitations. The selected indicators are highly dependent on data which ignore the industry background and the influence of time. Also, the function of parameter adjustment in RapidMiner needs to be further optimized.

Acknowledgement

This research was supported by Science Foundation of Beijing Language and Culture University –the Fundamental Research Funds for the Central University (20YJ040001,17PT05).

References

[1] Benjamin Graham, 2013, Securities Analysis [M]. David L. Dodd. The Sixth Edition. (China: China Renmin University Press)
[2] Zhao Ming, 2018, Research on the Investment Value of Equipment Manufacturing Enterprises Based on Machine Learning [D] (Jilin: Jilin University)
[3] Xia Xin, 2011, Research on the Construction and Application of the Evaluation Index System of Venture Capital Value in China’s Media Industry [D] (Shanghai: Fudan University)
[4] http://www.scholarpedia.org/article/Ensemble_learning, 2019-5-4
[5] https://www.cnblogs.com/bnuvincent/p/7617853.html, 2019-5-5