Information System Construction and Research on Preference of Model by Multi-Class Decision Tree Regression

Zhanpeng Mao¹, *, Tingting Zheng², Zongqiang Lian²

¹School of Economics and Management, Changchun University of Technology, Changchun, Jilin, 130000
²College of Economics, Shenzhen University, Guangdong, 518000

*Corresponding author: zhanpengmao@ccut.edu.cn

Abstract. Through the construction of customer asset allocation decision preference model based on machine learning, the input variables are set as demographic variables, family economic conditions, personality psychological characteristics and risk attitude, and the output variables are customer asset allocation decision preference choices. Machine learning algorithms such as decision tree and support vector machine are used to predict customers' asset allocation decision preference, and compared with traditional prediction methods. The results show that the machine learning algorithm can predict customers' asset allocation decision preference to a certain extent, and its performance is more effective than the traditional prediction method. Combined with the sample data, this paper analyzes the impact of venture capital on the growth ability of small and medium-sized enterprises, uses entropy method to evaluate corporate growth ability and corporate governance respectively, and further establishes a regression model to verify the intermediary role of corporate governance. The structured macro-prudential monetary policy rules with different response coefficients to the leverage ratio of different enterprises can effectively improve the level of social welfare, and the combination of optimal response coefficients is different under different kinds of shocks. Venture capital involvement in small and medium-sized enterprises can significantly improve the growth ability of small and medium-sized enterprises, and the support of venture capital can help enterprises improve the level of corporate governance. The experimental results show that the scorecard model constructed by the proposed method has good stability. The more obvious the improvement of the level of social welfare caused by the structural macroprudential monetary policy rules. The accurate prediction of asset allocation decision preference helps to improve customer decision-making efficiency and satisfaction, and reduce the labor costs of financial institutions.

Keywords: Machine learning; Investment model; Decision making; Entropy weight method.
1. Introduction
For venture capital institutions, investing in corporate innovation projects can get higher returns, but it is also accompanied by high risks [1]. The uncertainty of the growth of start-up enterprises and the non-disclosure of enterprise innovation project information urge venture capital institutions to actively participate in enterprise operation, improve the value of invested enterprises and increase their own investment returns by helping enterprises formulate reasonable strategies and provide high-quality value-added services [2].

At present, small and medium-sized enterprises have become the new main force of China's economy and are the key to maintain the vitality of China's economic growth. The importance of high quality development of small and medium-sized enterprises is self-evident. Despite the rapid growth of small and medium-sized enterprises in recent years, there are still many problems in their growth process, which restrict the high quality development of small and medium-sized enterprises. Corporate governance is the propellant for the process of rapid growth, high-quality growth and high-quality development of enterprises. For small and medium-sized enterprises in the growth stage, corporate governance issues should not be underestimated.

Relying on these limited indicators and experience has been far from being able to cope with the diversity and complexity of today's financial service product purchase decisions, resulting in customers unable to make better decisions. Although a large number of financial APPs have emerged, they still recommend financial products only based on risk assessment results [3], which cannot meet customers' needs for decision-making assistance in overall asset allocation. Therefore, how to provide the optimal asset allocation scheme automatically and quickly based on the intelligent algorithm and the limited information of customers has become extremely urgent, which plays a crucial role in promoting the transformation and upgrading and healthy development of China's wealth management industry.

2. Construction and implementation of logical regression algorithm

2.1. Logical regression algorithm
Logical regression algorithm is a common algorithm used to solve classification problems [4]. For the input data, when this data is greater than our threshold value, the output is 1, when it is less than the threshold value, the output is 0. The output variables of this model are always in the range of 0~1.

The assumed function of logistic regression model is:

\[ h_\theta(x) = g(\theta^T X) \]  

(1)

\( X \) represents the eigenvector, and \( g(z) \) represents the logical function. \( G(z) \) is a commonly used logical function S-shaped function[5], expressed as \( g(z) = \frac{1}{1 + e^{-z}} \). By combining these two formulas, the hypothesis of logistic regression model can be obtained. \( h_\theta(x) \) is used to calculate the probability that the output variable = 1 for a given input variable \( x \) based on the selected parameters, \( h_\theta(x) = P(y = 1 | x; \theta) \). For example, for input \( x \), \( h_\theta(x) = 0.65 \), indicating that there is a 65% probability that \( y \) is a positive class and a 35% probability that \( y \) is a negative class[6].

The relationship between \( h_\theta(x) \) and \( Cost(h_\theta(x), y) \) is shown in Fig. 1
Put it into the cost function of logistic regression to get:

\[
J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log \left(1 - h_\theta(x^{(i)})\right)]
\]  

(2)

After obtaining such a cost function, we use gradient descent algorithm to find the parameters that can minimize the cost function. Gradient descent method is to follow the direction of the fastest descent rate each time and reach the lowest point of the curve after many iterations. In order to find the smallest parameter \(\theta\) in Equation , a random value is initialized each time, and then the value of \(\theta\) is iteratively updated.

2.2. Entropy weight method

Data normalization processing positive indicators:

\[
X'_y = \frac{X_y - \min \{X_j\}}{\max \{X_j\} - \min \{X_j\}}
\]

(3)

Negative indicators:

\[
X'_y = \frac{\max \{X_j\} - X_y}{\max \{X_j\} - \min \{X_j\}}
\]

(4)

Calculate the proportion of the j index of i enterprise:

\[
Y'_i = \frac{X'_y}{\sum_{i=1}^{m} X'_y}
\]

(5)

The index information entropy was calculated:

\[
e_j = -k \sum_{i=1}^{m} Y'_i \cdot \ln Y'_i
\]

(6)

Calculate index weight:

\[
W_i = \frac{1 - e_j}{\sum_{j=1}^{m} (1 - e_j)} \text{ Where } DJ = 1 - e_j \text{ is the redundancy of information entropy.}
\]
Score of unidirectional index evaluation was calculated:

\[ S_{ij} = X_{ij} \cdot W_i \]  

(7)

Decision tree is a popular classification algorithm because of its relatively fast and easy solving process, simple and easy to understand. Support vector machine (SVM) is a supervised learning method that requires label data sets. Based on the structural risk minimization in machine learning theory, support vector machine can deal with two or more kinds of data, which realizes the Vapnik-Chervonenkis dimension theory and the structural risk minimization principle. In this study, three methods are used for calculation and verification. The first group uses the decision tree model to predict the customer asset allocation decision preference, the second group uses the support vector machine model to predict the customer asset allocation decision preference, and the third group uses the K-Means model to predict the customer asset allocation decision preference.

This study uses four indicators to evaluate the model: accuracy, KL divergence, F1Measure and hamming distance. The combination of these four indicators can effectively evaluate the model.

3. Construction and implementation of logical regression algorithm

The decision tree model, support vector machine model and K-means model were used to predict the customer asset allocation decision preference at 0.05 confidence level, and the selection behavior of 75 random samples in the customer set was predicted based on the estimated results, and the predicted results were compared with the actual selection behavior of customer asset allocation.

| Tab.1 Three models predict the results |
|--------------------------------------|
| Model                      | Accuracy | KL divergence | F1Measure |
| Decision tree model         | 0.743    | 0.229         | 0.658     |
| Support vector machine model| 0.719    | 0.227         | 0.601     |
| K-means model               | 0.301    | 0.376         | 0.371     |

(a) Tree Confusions Matrix
The column of the matrix represents the customer's actual asset allocation category, and the row represents the customer's asset allocation decision preference prediction category. The diagonal of the matrix is the number of assets classified. A, B, C, D and E are savings deposits, bonds, bank wealth management products, stocks and funds, respectively. The figure shows the confusion matrix obtained by decision tree model, support vector machine model and K-Means model. It can be seen from the chart that savings deposits are the choice of most customers, which is the same as the existing research results. It shows that the model predicts that the largest number of people choose stocks for asset allocation, which is contrary to the existing research results. Thus it can be seen that the prediction accuracy of machine learning is better than that of the traditional model.

4. KS curve and ROC curve
Sort the samples according to the prediction result of the learner (note, the probability value of the positive example, non-0/1 variable) in order from the largest to the smallest -- this is the order in which the cut-off points are selected. Select the cut-off points in order, and calculate TPR and FPR -- you can also select only n cut-off points, at positions 1/N, 2/N, 3/N, etc. The horizontal axis is the percentage of samples (the maximum is 100%); the vertical axis is TPR and FPR respectively, and KS curve can be obtained.

The ROC characteristic curve is also called the receptivity curve, which is named because the points on the curve reflect the same receptivity. They are all responses to the same signal stimulus, but the results are obtained under several different criteria. The receiver operating characteristic curve is a coordinate diagram composed of false positive probability (probability of negative case error) as horizontal axis and recall as vertical axis, and the curve drawn by different results drawn by subjects due to different judgment criteria under specific stimulus conditions.
The evaluation method of ROC curve is different from the traditional evaluation method. It does not need this restriction, but allows intermediate state according to the actual situation. The test results can be divided into several ordered categories, such as normal, roughly normal, suspicious, roughly abnormal and abnormal, and then statistical analysis can be carried out. Therefore, the ROC curve evaluation method is applicable to a wider range. The drawing steps of ROC are as follows: the samples are sorted (from large to small) according to the predicted results of the learner (i.e. the probability value of the positive example and non-0/1 variable) -- this is the order in which the cutoff points are selected; Then select the cut-off points in order, and calculate TPR and FPR. It can also select only n cut-off points, respectively in 1 / n, 2 / n, 3 / n and so on; Finally, all the points (TPR, FPR) are connected, that is, the ROC diagram. The ROC curve is shown in Figure 3.

Fig.3 ROC evaluation curve

### Tab.2 Statistical table of regression effect of each model

| Model   | F       | P     | R² | VIF(max) |
|---------|---------|-------|----|----------|
| Model 1 | 27.968  | 0.000 | 0.254 | 1.202    |
| Model 2 | 11.092  | 0.000 | 0.108 | 1.202    |
| Model 3 | 23.627  | 0.000 | 0.254 | 1.202    |

Model 1 takes VC as the explanatory variable, Growth as the explained variable, Age and Size as the control variable, and mainly studies whether venture capital can improve the growth ability of small and medium-sized enterprises. The collinear diagnosis of the main variables in the model shows that the regression result is reliable.

Model 2 takes VC as the explanatory variable, U as the explained variable, Age and Size as the control variable, and mainly studies whether venture capital participation can improve the level of corporate governance. According to the collinearity diagnosis of model 2, it is found that the maximum value of VIF is 1.202 less than 2, which indicates that there is no multicollinearity among the variables in model 2, so the regression result is reliable.

Model 3 takes VC and U as explanatory variables, Growth as explanatory variables, Age and Size as control variables, and mainly studies whether corporate governance plays an intermediary role between venture capital and corporate growth.

### 5. Conclusion

Venture capital involvement in small and medium-sized enterprises can significantly enhance the growth ability of enterprises. Venture capital has broadened the traditional financing channels and injected new vitality into the new financing channels.

Applying machine learning method to financial institutions to assist decision-making in customer asset allocation can help achieve reasonable allocation of customer assets, improve returns and reduce risks, which has practical significance for financial institutions to provide precise customer service, reduce business risks and improve business performance.
References

[1] Zhu Shaoying, Xu Yu, he Zhengwen. Multi-objective decision-making model and solution of venture capital under different risk preferences [J]. Systems Engineering, 2003,21 (003): 13-17.

[2] Huang Fuguang, Jia Ximeng, Tian Li. Knowledge background and high-tech investment preference of senior management team of venture capital institutions [J]. Management Science, 2016,29 (005): 31-44.

[3] concerned about such as, Li Yuanqin. Research on Venture Capital preference [J]. Accounting Newsletter: medium, 2016.

[4] Peng Weimin. The choice of venture capital behavior of investors [J]. Journal of Southwest Jiaotong University, 2005,40 (6).

[5] Chen Yiyang, Ma Binhua. Research on the influence of AB Stock Market investors' risk preference on Venture Capital decision-making [J]. Financial News, 2019000 (002): 175-176.

[6] Ma Yan, Jiang Tao. Equity preference of external financing of Chinese enterprises [J]. Northern Finance, 2010 (7): 15-18