Application of Machine Learning in Enterprise Growth Assessment: A Study of China's A-Share Listed Enterprises in Multiple Industries from 2000 to 2022

Chunxue Yao[*]

Statistics and Mathematics College, Inner Mongolia University of Finance and Economics, Hohhot, China
**Corresponding author*

Abstract: *In the rapidly changing digital business environment, accurately assessing corporate growth is of utmost importance for corporate strategy formulation, investment decisions, and industry research. This study focuses on multiple industries, collects data from 84 Chinese A-share companies from 2000 to 2022, and uses time-series analysis to explore data trends. After data processing, the K-means clustering algorithm is adopted. The elbow method is used to determine the clustering strategy, and K-means classification is carried out. According to the growth characteristics, the enterprises are divided into three categories. Dimensionality reduction is performed to find 13 factors to assist in decision-making, showing the commonalities and differences in different stages. Researchers use the K-Nearest Neighbor algorithm (KNN), Classification and Regression Tree algorithm (CART), and Support Vector Machine (SVM) to build a growth prediction model. By optimizing with time-series data, the accuracy of the model has been significantly improved. The research results prove that the combination of machine learning and time-series analysis is accurate and reliable in evaluating and predicting corporate growth, and can effectively identify growth patterns. The results of this study are helpful for enterprises to formulate strategic plans, assist investors in making decisions, and provide references for policymakers to support industrial development.*

Keywords: *Time series, Machine learning, Multiple industries, Cluster analysis*

1. Introduction

In the current era where digitalization drives the transformation of the global economy, the importance of the Enterprise Total Growth (ETG) assessment has become increasingly prominent[1]. It provides crucial guidance for investors' decision - making and enterprises' strategic arrangements. Due to the booming development of China's A - share market from 2000 to 2022, the number of listed companies has increased dramatically. This study selects 84 industries, covering a wide range of fields such as finance, market, and operation. In order to deeply explore the internal driving factors and potential patterns of enterprise growth, this study selects the highly representative operating revenue growth rate (Growth), total asset growth rate (Asset Growth), and net profit growth rate (Net Profit Growth) as the explained variables. The operating revenue growth rate (Growth) directly reflects the expansion effect of an enterprise's products or services in the market and is a key indicator for measuring an enterprise's market competitiveness. The total asset growth rate (Asset Growth) reflects the expansion speed of an enterprise's asset scale and is related to strategic measures such as investment, mergers, and acquisitions. The net profit growth rate (Net Profit Growth) accurately presents the upward trend of an enterprise's profitability and is related to the foundation of an enterprise's sustainable development. The relevant code is available at https://github.com/chunxueyao/ETG.

Focusing on these explained variables and leveraging the powerful data processing capabilities of machine learning technology[2]. this study conducts an in-depth exploration of the correlation relationships between them and numerous independent variables, such as the fixed asset ratio (FIXED), gross profit margin (Gross Profit), investment level 1 (Invest1), investment level 2 (Invest2), and other indicators.

However, several fundamental research questions for ETG remain: (1) Given the complex and changeable business environment and significant differences in enterprises' internal structures and operations, accurately determining the number of clusters is crucial for comprehensively assessing ETG, evaluating enterprise development, and making rational plans. (2) Research China's A-share market from 2000 to 2022. Employ a time-series prediction model to forecast ETG and identify relevant patterns.

To address the research questions, this study makes two major contributions. 1.The elbow method is applied to 52,531 data points from 84 industries[3]. By analyzing the data distribution, the inflection point is accurately located to clarify the clustering strategy. Subsequently, the k - means algorithm is utilized to classify the data based on their internal relevance. The Euclidean distance is introduced to sort the clusters according to the distance between the cluster centers and the origin, and the variables are discretized to enhance the accuracy.2.Key features such as seasonal, long - term, and periodic fluctuations are extracted from the time - series data[4]. Traditional machine - learning models are employed to reduce the dimensionality of the feature variables in consideration of the time dimension. The most effective model is selected for dimensionality reduction, and 13 key factors are identified to predict the future of enterprises, thus providing support for management decision - making.

In the special research on enterprise development, the operating revenue growth rate (Growth), the total asset growth rate (Asset Growth), and the net profit growth rate (Net Profit Growth) are selected as the core explained variables. These indicators accurately reflect the development potential of enterprises from multiple dimensions such as market revenue, asset scale, and profit level, laying a solid foundation for subsequent research. This study aims to break through the limitations of traditional assessment methods and construct an accurate model. Through time - series prediction, it endeavors to reveal the growth trends, explore the potential, and assist enterprises and investors in seizing the initiative in the volatile market.

2. Literature Review

2.1 Assess Enterprise Total Growth (ETG)

From 2000 to 2022, the financial industry benefited from economic growth and the improvement of the financial system and achieved steady development[5]. The cyclical industries witnessed significant fluctuations along with economic fluctuations and policy adjustments[6]. The consumer industries performed well due to the increase in residents' income and stable demand. The technology industries rose rapidly with the help of technological revolutions. The infrastructure industries continued to develop driven by policy support and the urbanization process.

2.2 Predict Enterprise Total Growth (ETG)

Against the backdrop of the "dual carbon" goals, traditional cyclical industries such as steel and non-ferrous metals are also undergoing a green transformation, which indicates that green and sustainable development has become an inevitable trend in industrial development[7]. The technology industry has risen rapidly thanks to the technological revolution, demonstrating that innovation and technological upgrading are the core driving forces for industrial development. All industries should increase their investment in research and development and promote technological innovation to enhance their competitiveness and adapt to market changes. The consumer industry has performed well due to the increase in residents' income and stable demand. Moreover, with the upgrading of consumption, consumers' demands for product quality, personalization and so on are constantly rising[8]. Enterprises need to have an in-depth understanding of the changing trends of consumer demands and adjust their products and services in a timely manner to meet the market demands.

3. Methodology

3.1 Data collection

This study collected relevant data of domestic listed A-share companies from the data set of China Stock Market Accounting Research (CSMAR). Considering the unique nature of the bankruptcy process in the financial sector, the analysis focuses on the time series data of companies listed in the A-share market from 2000 to 2022, with each company represented by one data point per year.

3.2 Data processing

In the data processing stage, this study first preprocessed the data. It removed observations with missing variable values and interpolated for the missing ones. Then, it excluded listed companies with a listing duration of less than five years.

For variable selection, the study used correlation analysis with the Pearson correlation coefficient method for continuous variables. The coefficient ranges from -1 to 1, indicating the degree of correlation. Low - correlated explanatory variables were removed.

After calculating the correlation coefficient and eliminating outliers, 13 variable names were chosen as explanatory variables, such as 'Gross Profit', 'FIXD', 'Invest1', etc. Next, these variables will be used to show their relationships with the dependent variables ('Growth', 'Asset Growth', 'Net Profit Growth') and construct predictive models. A table named "Table.1. Definition and Description of Attributes" will be drawn to explain these variables. A table will be drawn to explain these variables.(Del.)

Table.1. Definition and Description of Attributes

Attribute	Explanation of attributes	Attribute Name	Attribute Description
Gross Profit	Gross margin on sales	X1	(Operating Income - Operating Costs) / Operating Income
FIXED	Proportion of fixed assets	X2	Net Fixed Assets / Total Assets
Invest1	Level of investment	X3	Construct cash/opening total assets paid for fixed assets, intangible assets, and other long-term assets
Invest2	Level of investment	X4	(Cash paid to construct fixed assets, intangible assets and other long-term assets + net cash paid by subsidiaries and other business units) / Total assets at the beginning of the period
Dual	The two positions are one	X5	The chairman and the general manager are the same person as 1, otherwise it is 0
Top1	The shareholding ratio of the largest shareholder	X6	Number of shares held by the largest shareholder / total number of shares
ATO	Total Asset Turnover	X7	Operating Income / Average Total Assets
TMT Age	Average age of management	X8	The average age of directors, supervisors and senior executives
INST	Percentage of shares held by institutional investors	X9	Total number of shares held by institutional investors / total number of shares
WW	Financing constraints2	X10	The larger the WW index, the greater the financing constraints
Listed year	Year of listing	X11	Year of listing
Establish year	Year of establishment	X12	Year of establishment
Mid	Central Region	X13	1 for the central region, 0 otherwise

3.3 Elbow method for numbers of clustering

When determining the number of clusters using the elbow method, data preparation is carried out first. For the explanatory variables, specifically the columns "Growth", "Asset Growth", and "Net Profit Growth", only the IDs are retained for which the values in these three columns are all greater than - 15 and less than 15. The data within this range is relatively abundant and concentrated, making it easier to perform clustering.

Subsequently, based on the preliminary understanding of the data and experience, the range of the number of clusters is determined. In this case, the selected range is from 6 to 10. For each value of the number of clusters K within this range, the K - Means algorithm is applied to partition the data into K different clusters, with each cluster containing a certain number of data points.

Then, for the clustering results of each K value, the within - cluster sum of squares (WCSS) is calculated. The smaller the sum of the squares of the distances from each data point to the centroid of its corresponding cluster, the closer the data points within the cluster are, indicating a better clustering effect where K is the number of clusters in clustering. C_i represents the i cluster.

$$\mathrm{WCSS} = \sum_{i=1}^{K} \sum_{X_j = C_i} d(X_J, \mu_I)^2, \quad (i=1,2,3......,k), \tag{1}$$

Subsequently, a line chart is plotted with the number of clusters K as the abscissa and the corresponding WCSS value as the ordinate. As K increases, the WCSS value generally decreases gradually. This is because more cluster centers can fit the data better, making the data points within the clusters more closely grouped.

Finally, by observing the elbow plot, the turning point where the curve changes from a rapid decline to a slow decline is sought. The K value corresponding to this "elbow" point is the optimal number of clusters determined by the elbow method. As shown in the Figure 1, it is K = 3. After this point, increasing the number of clusters contributes less to the decrease in the WCSS value, and further increasing the number of clusters may complicate the model or lead to over - fitting.

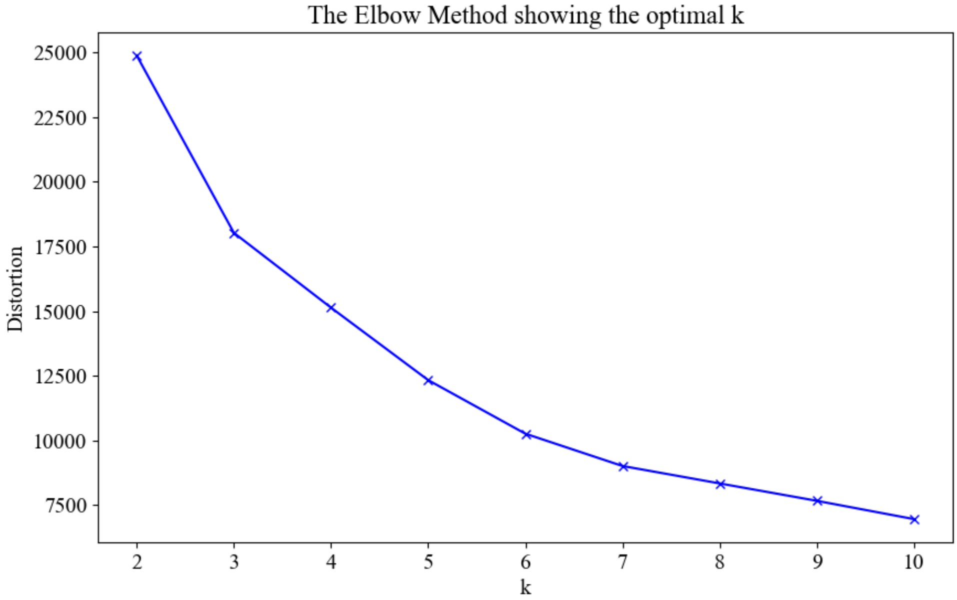

Figure 1. The elbow method showing the optimal k

3.4 Based on clustering for assessing enterprise total growth (ETG)

This paper uses the K - means clustering model to cluster the data in two major sample sets. With this model, the K value can be flexibly changed to determine the clustering scheme. To implement K - means clustering, first, the distance between two samples needs to be determined. Then, appropriate cluster centers are selected, and dynamic programming is carried out based on the distance. A function is determined to evaluate the clustering effect, and a clustering center scheme with good effect and relatively small data volume is selected. In this paper, the data with values between - 15 and 15 in "Growth", "Asset Growth", and "Net Profit Growth" are selected for clustering.

Because there are significant differences in the original data, in order to reduce such differences, this paper first normalizes each data index. For the i sample in the j sample set, the normalization is carried out, and the formula is obtained as follows:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \ , \tag{2}$$

where X is the original data, $X_{min}$is the minimum value in the data set, $X_{max}$is the maximum value in the data set, and$X_{norm}$is the normalized data. The value range of the obtained normalized index is[0, 1].

K cluster centers need to be selected. The sample closest to the center of the value space is taken as the first cluster center. To avoid the concentration of cluster centers, this paper uses the determined cluster centers to search for the next cluster center until K cluster centers are found. This can ensure the effectiveness of clustering.where X and Y are two points in an n - dimensional space.

$$d((X_J, \mu_I) = \sqrt{\sum_{l=1}^{n}(X_{jl} - \mu_{il})^2}, \tag{3}$$

In this study, the elbow method clustered A - share enterprises into three categories by overall growth variables. This balances intra - cluster compactness and inter - cluster differences, identifying three enterprise groups with distinct growth rates.

Industry - based cluster analysis of the A - share market shows industry - specific enterprise growth levels. Enterprises in the same cluster have similar growth patterns and drivers. Studying these three types helps determine growth model drivers and assess market health. More high - growth enterprises suggest a vibrant market; more low - growth ones imply challenges.

Yet, this method has limits. With numerous A - share enterprises, data collection and processing can be error - prone, risking misclassification. Still, cluster analysis aids investors in target screening and helps enterprises formulate strategic plans.

3.5 Based on machine learning for predicting enterprise total growth (ETG)

K - Nearest Neighbors (KNN) Algorithm: The K - Nearest Neighbors (KNN) algorithm falls into the category of supervised learning. It is a common and fundamental method in classification and regression tasks. Different from other algorithms that derive an explicit model from the training data, KNN does not directly generate such a model. Its prediction process is unique. When encountering a new sample, it measures the distance between this sample and each sample in the training set. Based on this distance information, it selects the K training samples that are closest. Finally, it uses the information contained in these K samples to complete the prediction.

CART Decision Tree: The CART (Classification and Regression Tree) decision tree has a remarkable feature. Whether dealing with classification or regression problems, it will generate a binary tree structure, that is, each node will split into two child nodes. When used for classification problems, the CART algorithm constructs a classification tree. In this process, it adopts the Gini impurity as the default splitting criterion. It represents the proportion of samples belonging to class j in the dataset, and k is the total number of classes.

$$\mathrm{Gini}(p) = 1 - \sum_{j=1}^{k} p_j^2, \tag{4}$$

Specifically, when calculating, it reflects the probability of randomly selecting two samples from the dataset with different class labels.

The Support Vector Machine (SVM):SVM is a powerful supervised - learning model that covers tasks including classification, regression, and outlier detection. Its core idea is to determine an optimal hyperplane (decision boundary) which can maximize the margin (or interval) between different classes. To handle complex non - linear relationships, the SVM introduces the Radial Basis Function (RBF) kernel and balanced class weights,where$\boldsymbol{\gamma}$ is the coefficient of the kernel function, and $\|\boldsymbol{x} - \boldsymbol{x}'\|$ represents the Euclidean distance between two samples.

$$K(x, x') = \sqrt{-\gamma\|x - x'\|^2}, \tag{5}$$

Specifically, the Radial Basis Function can map features into an infinite - dimensional space, providing a powerful tool for dealing with complex non - linear relationships. This mapping ability gives the SVM a strong advantage when dealing with non - linearly separable data. It can find a more suitable decision boundary, improving the accuracy of classification and regression. However, accordingly, its parameter adjustment and computational complexity may be relatively high, requiring certain skills and experience for optimization.

4. Results

4.1 Results of assessing enterprise total growth (ETG)

After repeated trials, it was determined that the number of categories for "Growth", "Asset Growth", and "Net Profit Growth" in the enterprise's overall growth is 3. The clustering results are shown in the table, which basically meet the expectations, as shown in Table 2.

Table 2. Specific Classification Results Table of Enterprise Total Growth

Type	Sample point serial number
0	11,20,23,25,2636,55,158...
1	6,152,608,619,632,909,917...
2	886,915,910,919,920,926...

The cluster centers corresponding to the overall growth of enterprises are shown in the following table, as shown in Table 3.

Table 3. Coordinate Table of Cluster Centers for Enterprise Total Growth

Categories	Growth	Asset Growth	Net Profit Growth
0	0.15	0.25	0.35
1	0.55	0.65	0.75
2	0.85	0.95	1.05

4.2 Results of predicting enterprise total growth (ETG)

We analyze the overall growth of enterprises from four aspects:

Accuracy: It's the proportion of correctly predicted samples. The SVM model tops with 94.39% accuracy in A - share enterprise growth forecasting. KNN follows closely at 94.08%, while CART has a lower 89.04%.

Precision: This is the proportion of truly positive samples among those predicted positive. CART leads with 90.10%, with KNN at 89.08% and SVM at 89.09%.

Recall: It shows the model's ability to identify all positive samples. SVM and KNN have higher recall rates, 94.39% and 94.08% respectively, compared to CART's 89.04%.

F1 - score: A combined metric of precision and recall. SVM scores 91.67%, higher than KNN's 91.51% and CART's 89.56%, showing its better balance.

Overall, for analyzing A - share enterprise growth, the SVM model performs best among these evaluation metrics, offering more accurate classification and prediction, as shown in Table **4**.

Table.4. Performance indicators of overall corporate growth

Model	Accuracy	Precision	Recall	F1 Score
KNN	94.08%	89.08%	94.08%	91.51%
CART	89.04%	90.10%	89.04%	89.56%
SVM	94.39%	89.09%	94.39%	91.67%

5. Conclusions

During this research, precisely determining cluster numbers was crucial for enterprise growth assessment. Given the complex business setting and varied enterprise structures, this was essential for development evaluation and planning. We also studied the 2000 - 2022 Chinese A - share market with a time - series model to predict growth and uncover patterns.For variable selection, the Pearson correlation method helped us sift variables and address challenges. Using the K - Means algorithm, we classified enterprises by growth features. Then, with KNN, CART, and SVM, we built a model with 13 explanatory and 3 explained variables to assess growth trends.Through strict procedures, we obtained a highly accurate and reliable model. Clustering revealed distinct growth traits among enterprises. In model evaluation, SVM led in accuracy (94.39%) and F1 - score (91.67%), CART in precision (90.10%), and SVM and KNN in recall (94.39% and 94.08% respectively). Overall, SVM achieved a good balance.Our research offered a scientific and efficient approach for enterprise growth assessment.Our research provides a scientific and efficient method for enterprise growth assessment. It is beneficial for enterprise strategy formulation, investor decision - making, and industry policy - making, and promotes the economic environment to develop towards prosperity and stability. In the future, we expect to deeply integrate big data, artificial intelligence, and machine learning technologies to continuously iterate and optimize the model. This approach not only broadens its application in emerging industries and multinational enterprises but also helps with enterprise risk management and precise resource allocation, injecting a continuous stream of new impetus into the robust development of the global economy in all aspects..

References

[1] Yu L, Gao X, Lyu J, et al. Green growth and environmental sustainability in China: the role of environmental taxes[J]. Environmental Science and Pollution Research, 2023, 30(9): 22702-22711.

[2] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

[3] Cui M. Introduction to the k-means clustering algorithm based on the elbow method[J]. Accounting, Auditing and Finance, 2020, 1(1): 5-8.

[4] Wang X, Smith K, Hyndman R. Characteristic-based clustering for time series data[J]. Data mining and knowledge Discovery, 2006, 13: 335-364.

[5] Wang R, Tan J. Exploring the coupling and forecasting of financial development, technological innovation, and economic growth[J]. Technological Forecasting and Social Change, 2021, 163: 120466.

[6] Li C, Zhang Y, Li X, et al. Artificial intelligence, household financial fragility and energy resources consumption: Impacts of digital disruption from a demand-based perspective[J]. Resources Policy, 2024, 88: 104469.

[7] Wang J, Liu W, Chen L, et al. Analysis of China's non-ferrous metals industry's path to peak carbon: A whole life cycle industry chain based on copper[J]. Science of The Total Environment, 2023, 892: 164454.

[8] Teepapal T. AI-driven personalization: Unraveling consumer perceptions in social media engagement[J]. Computers in Human Behavior, 2025, 165: 108549.