Research in Economics and Management
ISSN 2470-4407 (Print) ISSN 2470-4393 (Online)
Vol. 11, No. 1, 2026
www.scholink.org/ojs/index.php/rem

Original Paper

Key Associated Characteristics and Industry Heterogeneity of Corporate Data Asset Potential in China-Evidence from Generative Models and Explainable Machine Learning

Junhong Luo[1], Huijuan Zheng[1], & Junxi Huang[1]

[1] Guangdong University of Finance and Economics, Guangzhou, 510320, China

Received: April 3, 2026 Accepted: April 20, 2026 Online Published: May 4, 2026

doi:10.22158/rem.v11n1p69 URL: http://dx.doi.org/10.22158/rem.v11n1p69

Abstract

This study takes A-share listed firms that recognized data resources as assets in their 2024 annual reports as the initial sample. To address the issue of limited sample size, a generative model is adopted for data augmentation. Based on the augmented data, an interpretable machine learning framework is developed to explore the main firm characteristics associated with data asset potential, as well as the nonlinear patterns underlying these associations. The results show that innovation investment and firm size are the two most relevant characteristics. Their marginal contributions to data asset potential exhibit clear nonlinear threshold effects. In contrast, conventional profitability metrics have weak explanatory power, and their influence is contingent on the level of innovation input. A subsequent analysis covering the entire market reveals that the information technology sector and the scientific research and technical services sector tend to have substantially higher data asset potential. This study not only offers new empirical evidence on the correlates of data asset potential but also provides practical insights for the design of data factor market institutions, asset valuation practices, investment decisions, and corporate strategic planning.

Keywords

data asset, generative model, lightgbm, shap, strategic information systems

1. Introduction

1.1 Purpose of the Study

The systematic cultivation of data element market and the mining of potential data value have become the key objectives in China's strategic policy pattern. Landmark documents, including "opinions on establishing a more perfect market-oriented allocation mechanism of production factors" and "on

building a data-based system", have clearly promoted the transformation to the practice mode of data asset capitalization. In response to these requirements, the Ministry of Finance issued the Interim Provisions on accounting treatment of enterprise data resources in August 2023. These regulations will take effect on January 1st, 2024, marking a watershed moment for China's Digital Economy: the national accounting framework provides a clear agreement for the first time to recognize the attributes of data assets and regulate information disclosure. This change effectively bridged the gap between the long-term theoretical discussion and rigorous accounting application.

2024 is the first year for the implementation of these standards. The capitalization behavior of A-share listed companies provides an important sample for evaluating the effect of policies and understanding the changes in market sentiment. The empirical monitoring of the asset appraisal institute of Capital University of economics and business shows that as of April 30, 2025, only 1.85% of enterprises (100 of 5401) have officially included data resource capitalization in their 2024 annual report. Although this low adoption rate indicates that data capitalization is still in the initial exploration stage, the actions of these 100 "pioneer" enterprises provide indispensable intelligence for revealing the value of off balance sheet data that was previously difficult to detect. For market participants, the main analysis obstacle is to identify those enterprises that have huge data potential but have not been formally reflected in the financial statements. Therefore, this study investigates the measurable signals shared by early adopters to determine whether these commonalities can reliably predict the potential of data assets in the broader market.

The existing literature provides a basic understanding of data assets, but fails to solve the above identification challenges. Existing empirical studies rely heavily on keyword based text analysis to construct proxy variables, which often overemphasizes economic results at the expense of understanding the internal composition or nonlinear dynamics of data assets. In addition, the universality of direct case or event studies on capitalization practice is often limited due to the inherent characteristic of scarce samples. In order to make up for these deficiencies, this study proposes a diagnostic framework centered on "data asset potential", which is defined as the intrinsic economic value of data resources that are not recognized on the corporate balance sheet at present. We have developed a hybrid method that combines data enhancement and interpretable machine learning in three stages: (1) a more robust sample size is synthesized from the initial data using a table variation autoencoder (Tvae); (2) The diagnostic model was constructed by lightgbm and shap to ensure the accuracy and interpretability of prediction; (3) Systematically map the characteristic drivers and industry heterogeneity of data asset potential.

1.2 Contributions

This study has three main contributions. First, method innovation. This study bypasses the two obstacles of sample scarcity and complex value correlation model, which is the characteristic of the early stage of data capitalization. By integrating a framework of data expansion and interpretable machine learning, we provide a feasible technical route for the identification of off balance sheet data assets. Second, empirical depth. In addition to determining that innovation investment and organizational size are the main drivers

of data asset potential, this study also uses shap analysis to clarify its potential nonlinear dynamics and interaction effects. In addition, the study reveals obvious industry level heterogeneity, provides new empirical evidence, and improves our understanding of how data value is expressed in different industries. Third, practical significance. The diagnostic framework and the resulting evidence provide strategic benchmarks for a wide range of stakeholders. Specifically, these findings help regulators improve the disclosure agreement, enable evaluators to innovate their evaluation perspective, enable investors to discover potential data value, and provide a roadmap for enterprises to promote their data-driven strategy.

1.3 Literature Review

The research of data assets has experienced a process from basic theory to practical exploration. Data has the characteristics of non competitiveness, low marginal replication cost, value volatility and context dependence (Jones & Tonetti, 2020). Compared with traditional assets, realizing data value requires different ways. Xu Xianchun et al. (2022) proposed a methodological framework for the recognition and measurement of enterprise data assets from the perspective of national economic accounting. Luo Mei et al. (2023) proposed a methodological framework for the recognition and measurement of enterprise data assets from the perspective of accounting. With the implementation of the Interim Provisions on accounting treatment of enterprise data resources, the capitalization of data assets has moved from theory to practice. Through case studies, relevant studies have explored practical challenges and strategies in data ownership, governance, recognition of capitalized amounts, and subsequent measurement (Zhou Yarong et al., 2025; Ren Mudan et al., 2025).

With the development of data capitalization practice, its economic consequences have become the focus of empirical research. The existing literature mainly uses two methods: event research and text analysis. The event study examines the short-term market response to data asset capitalization events. For example, m.schreieck et al. (2024) conducted an event study based on the digital platform strategy announcement in a global sample of 165 banks. They reveal the complex interaction between the strategies for artificial intelligence and digital platforms and the returns of bank stocks in different markets around the world. Chenzhibin, Yang Jing (2025), Huang can and others (2024) found that the promulgation of the Interim Provisions on accounting treatment of enterprise data resources sent a positive signal to the market and significantly improved the enterprise value, especially in data intensive enterprises. Zhang Junrui et al. (2024) analyzed the interim report of 2024 and found that companies that disclosed the actual data asset amount received more positive market reaction than companies that only disclosed keywords. These studies provide direct evidence of the impact of data assets on the economy, but they focus on short-term market signals and are limited by small sample size, so it is difficult to explore complex potential mechanisms. On the other hand, some studies use text analysis to quantify the degree of enterprise digital transformation. By constructing proxy variables of data assets, they link data assets with a series of economic consequences, and confirm the impact of digital transformation on enterprise performance. M. Zareie et al. (2024) used text analysis to measure the digital transformation of enterprises by the frequency of digital terms in the enterprise 10-K report, and found that the digital transformation score

(DGS) was related to enterprise value. Importantly, organizational capital, corporate governance quality, information quality, and the era of corporate IPO play an important role in the value creation of enterprise digital transformation.

In China, data asset disclosure improves the efficiency of information supply and dissemination, reduces the synchronization of stock prices, and improves pricing efficiency (lishigang et al., 2025). In enterprise financing, data capitalization eases the financing constraints of "specialized, refined, featured and innovative" small and medium-sized enterprises (Heying et al., 2024), and significantly reduces the cost of debt financing (niubiao et al., 2024). In terms of innovation and development, data assets promote innovation investment by easing financing constraints and promoting R&D Cooperation (Li Jian et al., 2023; Mao Chunmei et al., 2024). Although these studies reveal the extensive economic effects of data assets using large data sets, its core limitation lies in the accuracy of proxy variables. The indicators based on word frequency may not accurately reflect the real data stock of the company, and may be affected by strategic disclosure behavior, which limits its ability to capture the real level of data capitalization, handle non-linear relationships and high-dimensional data. In view of the above limitations, the latest progress of machine learning has significantly enhanced data modeling, prediction performance and model interpretability across different domains. Xu et al. (2019) proposed ctgan, a condition generation countermeasure network tailored for synthetic table data. The heterogeneity of tabular data (e.g., mixed numerical and categorical variables) and privacy constraints pose challenges for direct modeling or sharing. Ctgan solves these problems by introducing the condition vector to model the unbalanced classification distribution, and using pacgan to enhance the diversity of samples through classification. The framework optimizes the joint modeling of numerical and classification features, and generates synthetic data closely related to the statistical characteristics of real data. Ke et al. (2017) introduced lightgbm, a high-performance gradient lifting decision tree (gbdt) framework designed to solve the computational bottleneck of large-scale high-dimensional data sets. Lightgbm is superior to traditional gbdt frameworks in speed and accuracy, such as xgboost, especially for high-dimensional and sparse tabular data. Lightgbm complements ctgan as a downstream model for evaluating the quality of synthetic data, and ctgan's synthetic data can expand lightgbm's training set in the case of data scarcity. Gu et al. (2020) discussed the application of machine learning in financial asset pricing, and solved the limitations of traditional models such as CAPM and Fama French, which rely on linear assumptions and predefined risk factors. Using neural network, random forest and gradient lifting techniques, the author models the U.S. stock return to predict the performance of a single stock and identify new risk factors. Their framework captures nonlinear relationships and high-dimensional feature interactions, and is superior to the traditional model in the prediction of out of sample returns. Lundberg and Lee (2017) introduced Shapley additive interpretation, a unified framework for interpreting machine learning model predictions. Shap uses the Shapley value of game theory to assign consistent and fair attribution scores to each feature, and quantify its contribution to the model output. Yoshioka's hypothetical evaluation of intangible fixed assets using generative artificial intelligence and machine learning introduces new applications of

generative artificial intelligence and machine learning to evaluate intangible assets (such as patents and trademarks), uses ctgan and other technologies to expand data, uses lightgbm and other models to make predictive evaluation, and potentially uses shap to explain the evaluation drivers (2024). Lu Yao et al. (2020) used boosting regression tree algorithm to test the predictive ability of multidimensional execution characteristics on enterprise performance, and achieved more accurate results than the traditional model. Heying et al. (2024) used lightgbm and shake to study the predictive ability of corporate and executive characteristics on corporate misconduct. Li Jianwei et al. (2025) also used lightgbm to discuss the factors affecting digital transformation and their relative importance. These papers jointly promote the application of machine learning in tabular data generation, efficient modeling, financial asset pricing, model interpretability and intangible asset evaluation. Ctgan's synthetic data can expand the training set of lightgbm and asset pricing model, and solve the data limitations in the financial and intangible assets appraisal tasks. The efficiency of lightgbm makes it suitable for evaluating the comprehensive data quality of ctgan and processing complex financial data sets, including those used for intangible assets evaluation. The framework of Gu et al. Benefits from the scalability of lightgbm and the interpretability of the new risk factors revealed by shap, which can be extended to the valuation of intangible assets by identifying key valuation drivers. Yoshioka's method may integrate generating AI (such as ctgan) to simulate intangible asset data and machine learning for evaluation. Shap provides transparency for feature contribution. Shap enhances all of these frameworks by providing insight into model behavior, ensuring trust in applications ranging from synthetic data generation to financial and intangible asset modeling. In short, the existing literature provides an important basis for the research of data assets, and emphasizes the direction of further exploration. Empirically, the research provides evidence of the value relevance of data assets, mainly focusing on the economic consequences of data capitalization. Whether it is through text analysis to investigate the impact on enterprise value or financing cost, or through event research to evaluate the short-term market reaction of capitalization announcement, the focus is on the external economic effect. However, the existing research rarely discusses the identification, composition and correlation characteristics of off balance sheet data assets. These research gaps pose two core challenges for this study: first, the practice of Chinese enterprises has only a one-year sample, and only 100 enterprises have included the data in the balance sheet. Training complex machine learning models on such a small sample, such as lightgbm, may lead to over fitting and unreliable conclusions. Secondly, the potential non-linear and interactive relationship between data asset related features is beyond the scope of linear econometric model. In order to meet the above challenges, this study combines shap with the framework of Gu et al. And Yoshioka et al. To build an analysis framework integrating data expansion and interpretable machine learning, systematically identify the key correlation characteristics of the asset potential of small sample data, explore the off balance sheet data asset potential of non capitalized enterprises, and test the heterogeneity of industries, thus providing an innovative research paradigm for data asset evaluation and empirical research.

2. Research Design

2.1 Data Sources, Variable Definitions, and Descriptive Statistics

The empirical analysis of this study centers on Chinese A-share listed firms in 2024. The initial seed sample comprises 100 listed companies that, as of April 30, 2025, had explicitly disclosed the capitalization of data resources in their 2024 annual reports. All firm-level data are obtained from the CSMAR database and the annual reports themselves. To maintain data completeness, any missing values in the sample are handled using median imputation. A detailed list of variable definitions and the corresponding calculation methods is provided in Table 1.

Table 1. Definitions and Descriptions of Main Variables

Variable Category	Variable Symbol	Variable Name	Calculation Method and Description
Dependent Variable	Data	Capitalized Data Asset	Total book value of data resources disclosed in annual reports, aggregated from the "Data Resources" sub-items under "Inventory," "Intangible Assets," and "Development Expenditure" in the balance sheet, per the "Provisional Regulations on the Accounting Treatment of Enterprise Data Resources".
Independent Variables	Size	Total Assets	Year-end value of "Total Assets" in the balance sheet.
	Revenue	Operating Revenue	Total revenue from main operating activities.
	Turnover	Total Asset Turnover	Operating revenue / Average total assets.
	Capex	Capital Expenditure	Cash paid for acquiring fixed assets, intangible assets, and other long-term assets.
	OCF	Net Operating Cash Flow	Net cash flow from operating activities (cash inflows minus outflows).
	Growth	Operating Revenue Growth Rate	(Current period operating revenue – Previous period operating revenue)/ Previous period operating revenue.
	ROA	Return on Assets	Net profit / Year-end total assets.
	ROIC	Return on Invested Capital	(Net profit + Financial expenses) / Average invested capital.
	GPM	Gross Profit Margin	(Operating revenue – Operating costs) / Operating revenue.
	TobinQ	Tobin's Q	Market value / (Total assets – Net intangible assets – Net goodwill).
	TechRatio	Technical Personnel Ratio	Proportion of employees with professional technical titles to total employees.

Variable Category	Variable Symbol	Variable Name	Calculation Method and Description
	RDRatio	R&D Investment Ratio	Total R&D investment / Operating revenue.
	RDPsnRatio	R&D Personnel Ratio	Number of R&D personnel / Total employees.
	Leverage	Debt-to-Asset Ratio	Total liabilities / Total assets.

Table 2 presents the descriptive statistics for the 100 seed sample firms after missing values were imputed. The average capitalization amount of data assets is about 34.8 million yuan. However, this variable shows wide dispersion and strong right deviation: the standard deviation is far greater than the average, the median is far lower than the average, and the gap between the maximum and minimum is large. This shows that in the early stage of capitalization, the size of capitalized data assets of different enterprises varies greatly, and the reported amount of a few leading enterprises is much higher than that of other enterprises. In addition, the sample shows significant differences in size and cash flow indicators, reflecting the actual differences between enterprises of different sizes. In terms of innovation investment, the average R&D investment rate is 7.071%, with a median of 3.84%, that is, more than half of the enterprises' R&D intensity is lower than the average level, showing a right deviation again. The average proportion of R&D personnel is 16.877%, indicating that the sample enterprises attach importance to human capital investment in innovation. The average ROA is close to zero and accompanied by a large standard deviation, indicating that profitability is widely dispersed in the whole sample.

Table 2. Descriptive Statistics of Seed Sample

Variable Symbol	N	Mean	Sd	Min	P50	Max
Data	100	34,796,998.333	1.203e+08	43,823.000	3,774,495.395	8.283e+08
Size	100	2.867e+11	1.224e+12	6.438e+08	2.033e+10	9.533e+12
Revenue	100	4.474e+10	1.241e+11	2.371e+08	6.819e+09	8.895e+11
Turnover	100	0.605	0.694	0.017	0.423	4.116
Capex	100	4.955e+09	1.950e+10	1,555,779.580	4.756e+08	1.560e+11
OCF	100	5.191e+09	4.649e+10	-2.048e+11	5.705e+08	3.157e+11
Growth	100	0.015	0.164	-0.546	0.018	0.418
ROA	100	0.009	0.058	-0.292	0.017	0.115
ROIC	100	0.019	0.078	-0.388	0.038	0.142
GPM	100	0.280	0.187	-0.118	0.261	0.932

TobinQ	100	2.031	1.951	0.742	1.477	14.171
TechRatio	100	0.289	0.239	0.004	0.201	0.872
RDRatio	100	7.071	8.984	0.060	3.840	43.000
RDPsnRatio	100	16.877	14.300	0.510	13.040	63.990
Leverage	100	0.505	0.212	0.058	0.508	0.925

2.2 Data Augmentation Using TVAE and Validation

This study uses a generative model for data augmentation, aiming not only to enlarge the sample size but also to capture and refine the underlying patterns present in the original data. More specifically, based on the 100 seed sample firms, we apply the TVAE (Tabular Variational Autoencoder) model from the open-source Synthetic Data Vault (SDV) library, which is designed for structured data (Xu et al., 2019), to generate a large augmented dataset of 100,000 records with high fidelity.

Tvae is a deep learning model based on the principle of variational automatic encoder. Its main advantage is that it can create new and consistent samples by learning and copying the complex joint probability distribution of the data set, rather than just copying the existing observations. Through the encoder network, Tvae compresses the high-dimensional data into the low-dimensional potential space that follows the specified probability distribution; Then, the decoder network reconstructs the new data from this potential space. This compression reconstruction process enables the model to extract the most essential and stable correlation structure between variables while avoiding the random noise existing in the original sample.

The implementation steps of Tvae in this study are as follows. First, we use the singletablemetadata module to check the seed data (after the median interpolation of the missing value) to ensure that the format and type of the generated data are consistent. Next, we initialize and train the tvasesynthesizer object. To ensure standardization and reproducibility, we use the default configuration parameters provided by the SDV library, as shown in Table 3. Finally, the trained synthesizer is used to generate augmented data sets for subsequent modeling.

Table 3. TVAE Synthesizer Default Configuration Parameters

Parameter	Default Value	Description
embedding_dim	128	Dimension of the latent space
compress_dims	(128, 128)	Dimensions of hidden layers in the encoder
decompress_dims	(128, 128)	Dimensions of hidden layers in the decoder
l2scale	1.00E-05	L2 regularization weight
batch_size	500	Batch size for each training iteration
epochs	300	Total number of training epochs

loss_factor	2	Balancing factor for reconstruction and KL divergence losses

In order to comprehensively evaluate the performance of the enhanced dataset, we intuitively compared the probability distributions of all variables. The comparison results are shown in Figure 1.

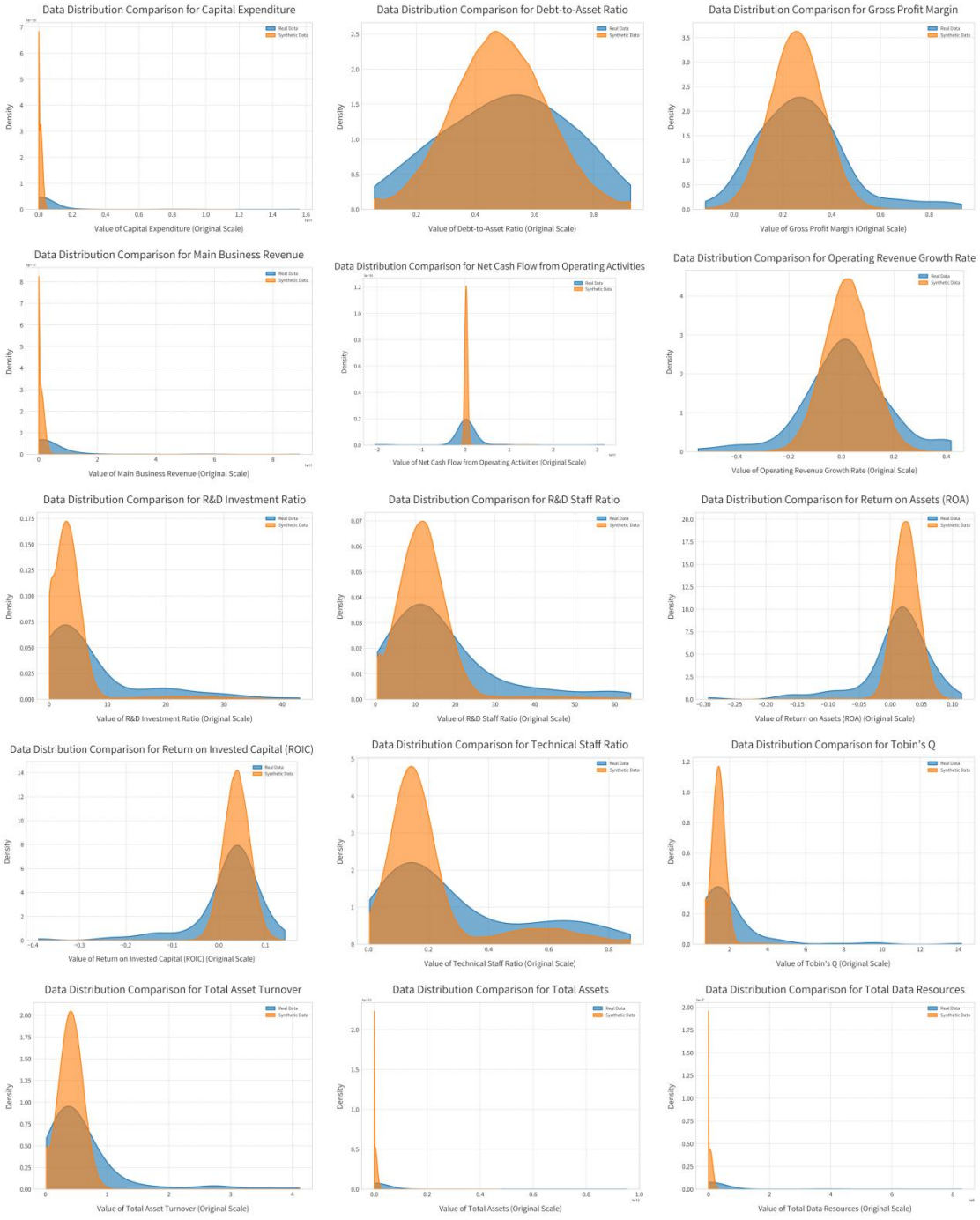

Figure 1. Probability Distribution Comparison of Core Variables between Seed Data and Augmented Dataset

Figure 1 compares the probability density distributions of key variables between the original seed data and the extended data set. Each subgraph corresponds to a key variable, where the x-axis represents the value of the variable and the y-axis represents the probability density. The blue filled area describes the actual data distribution of 100 seed companies, while the orange line shows the distribution of the enhanced data generated by the Tvae model.

The comparison shows that Tvae model successfully captures and reproduces the main statistical characteristics of seed data. First, whether the variable is a severely right biased scale measure, an approximately symmetrical rheological variable, or a typical ratio, the augmented dataset can well reflect the seed data in terms of distribution shape and central trend. The density curve of the enhanced data is in good agreement with the seed data in the peak position, skewness direction and data range, indicating that the model can accurately learn the edge distribution of each variable. Second, compared with the seed data, the extended data set shows a smoother and more centralized distribution, with a higher peak and a shorter tail. For example, in the case of return on assets (ROA), the seed data distribution is wider and flatter, while the extended distribution is similar to the standard bell curve. This result is expected because Tvae learns an idealized data generation process rather than simply reproducing random noise or extreme outliers in the original sample. Therefore, when training complex models on small samples, the enhanced data set helps to reduce the risk of over fitting, thus supporting the stable and reliable training of subsequent lightgbm models and providing a solid foundation for credible shape interpretation. To sum up, although there is a slight difference in the dispersion of distribution, the extended data set is consistent with the seed data in terms of core trend and structure. This can not only maintain the data fidelity, but also potentially improve the data quality through smoothing, making the enhanced data set very suitable for building a robust identification model of data asset potential.

2.3 Model Specification and Hyperparameter Optimization

Based on large-scale, high fidelity enhanced data sets, the data asset potential identification model is established by using lightgbm (lightgradient boosting machine) algorithm. Lightgbm is an effective framework based on gradient lifting decision tree (gbdt) method. Gbdt combines decision tree with ensemble learning, which has strong generalization ability, good training performance and robustness against over fitting, so it has been widely used in many fields. As an improved version of gbdt, lightgbm is developed and open source by Microsoft. It introduces gradient based one-sided sampling (Goss) and exclusive feature binding (EFB) technologies to achieve faster training speed and lower memory usage while maintaining high accuracy. These features make lightgbm particularly suitable for processing large-scale, high-dimensional data sets (Ke et al., 2017).

Regarding model specification and preprocessing, we first initialize the modeling environment using the setup function. This function automatically splits the augmented dataset into a training set (70%) and a test set (30%), and applies Z-score standardization to all numerical features. To ensure full reproducibility, a global random seed is set (session_id = 123). For hyperparameter tuning, we employ a random search strategy to optimize LightGBM's key hyperparameters. Specifically, the tune_model function is used to

perform a random search with 10 iterations within the predefined hyperparameter space of PyCaret, with the objective of maximizing the coefficient of determination (R^2). Each iteration is evaluated using 10-fold cross-validation to obtain reliable estimates of the model's generalization performance. The hyperparameter combination that yields the highest average R^2 across the cross-validation folds is selected, and the final model is trained on the full training set. The resulting optimal hyperparameter configuration is presented in Table 4.

Table 4. Final Hyperparameter Configuration for LightGBM Model

Hyperparameter	Optimal Value	Description
learning_rate	0.05	Learning rate
n_estimators	180	Number of base learners
num_leaves	10	Maximum number of leaf nodes
min_child_samples	96	Minimum number of samples per leaf
min_split_gain	0.7	Minimum split gain
feature_fraction	0.9	Feature sampling ratio
bagging_fraction	0.6	Data sampling ratio
bagging_freq	4	Data sampling frequency
reg_alpha	0.0001	L1 regularization coefficient
reg_lambda	0.1	L2 regularization coefficient
random_state	123	Random seed

2.4 Model Performance Evaluation

Then evaluate the final model on a separate test set, and the evaluation results are reported in Table 5.

Table 5. Performance Evaluation on Test Set

Evaluation Metric	Value	Description
Coefficient of Determination (R^2)	0.0461	Model goodness of fit
Root Mean Squared Error (RMSE)	6,648,811	Square root of the average squared difference between predicted and actual values
Mean Absolute Error (MAE)	5,490,950	Average absolute difference between predicted and actual values
Mean Squared Error (MSE)	4.42E+13	Average squared difference between predicted and actual values
Root Mean Squared Logarithmic Error (RMSLE)	2.8138	RMSE calculated after taking the logarithm of predicted and actual values

Evaluation Metric	Value	Description
Mean Absolute Percentage Error (MAPE)	42.95	Average percentage of absolute prediction errors relative to actual values

The R^2 of this model is 0.0461. It is worth emphasizing that the main goal of this study is not to develop a highly accurate prediction tool, but to build an interpretable diagnostic framework. When it comes to complex economic phenomena of data assets formed by many unobservable factors, the explanatory power (R^2) of models that rely only on public financial and operational data is essentially limited. As Gu et al. (2020) pointed out in their influential work of applying machine learning to asset pricing, asset returns are largely driven by unpredictable information, which makes it difficult to estimate the risk premium. Even if complex neural network models are used, the R^2 outside the sample for predicting monthly stock returns is only about 0.40%. In this case, low R^2 is not surprising; This actually shows that the model avoids serious over fitting of training data. More importantly, all error measures converge to stable finite values, which indicates that the model captures statistically significant associations that explain some changes in the potential of data assets, rather than simply performing random guesses. Therefore, the core value of the model is not its ability to make accurate numerical prediction, but its usefulness as a diagnostic framework for identifying off balance sheet data assets, so as to lay the foundation for subsequent in-depth shap analysis.

2.5 Model Interpretation Using SHAP

Although the evaluation results confirm the effectiveness of lightgbm model, its “black box” nature makes it difficult to directly understand its internal decision logic. In order to overcome this problem, we need a reliable tool to interpret the model, so that we can explore the criteria used by the model to identify the potential of data assets. The traditional global feature importance measurement may sometimes be inconsistent or misleading. Therefore, this study uses the Shapley addition (Shapley) for a deeper interpretation (Lundberg & Lee, 2017). Shap is an additive explanation model based on cooperative game theory, which is used to explain the output of tree based integrated machine learning method, and effectively alleviate the black box problem. The data asset potential of an enterprise can be regarded as the result of multiple contribution features. The Shapley value of each feature calculated by shap can better understand the potential mechanism. The main purpose of shap is to interpret a single prediction by quantifying the contribution of each feature. Specifically, for the predicted value of a given sample, the shap is expressed as an additive linear function: the sum of the baseline value plus the shap values of all features, as follows:

$$y_m = y_0 + \sum_{n=1}^{N} g\left(x_{mn}\right) \quad (1)$$

where x_{mn} is the (n)-th feature of the (m)-th sample,， y_m is the model's predicted value for the (m)-th sample, y_0 is the baseline prediction (typically the mean of all sample predictions), and (N) is the total number of features. $g(x_{mn})$ is the SHAP feature value of x_{mn}, representing its contribution to the prediction, calculated using the classic Shapley value formula:

$$g(x_{mn}) = \sum_{S \subseteq \{x_1, x_2, \ldots, x_N\} \setminus \{x_n\}} \frac{|S|!(N - |S| - 1)!}{N!} [f_x(S \cup \{x_n\}) - f_x(S)] \quad (2)$$

Where f_x is the trained LightGBM model, $\{x_1, x_2, \ldots, x_N\}$ is the set of all features, and S is any subset of features excluding feature x_n .This formula calculates the marginal contribution of feature (n) by averaging the change in model predictions when adding feature x_n across all possible feature combinations. Positive SHAP values indicate that the feature increases the predicted value, while negative values indicate a decrease.

In this study, we use SHAP values in two types of visualizations: summary plots and dependence plots. The summary plot aggregates SHAP values across all samples, allowing us to identify the most relevant features and the direction of their effects at the global level. In contrast, dependence plots provide a more detailed view, revealing the nonlinear and interactive relationships between the values of individual features and their marginal contributions to the model's output.

3. Empirical Results and Analysis

3.1 Identification Signals of Off-Balance-Sheet Data Assets: Feature Importance Analysis

In order to measure the contribution of each enterprise's characteristics to the potential of predicted data assets and understand the mode behind these contributions, we use the shap summary chart (Figure 2) to explain the trained lightgbm model in depth. This plot collects the shap values of all samples in the test set and sorts them according to the average absolute influence of characteristics on the model output. In the drawing, each line corresponds to a feature, and the global pattern of its influence is displayed in a decentralized distribution. Each point represents a single sample: its horizontal position (shap value) represents the marginal contribution of the feature to the prediction of the sample, and its color reflects the value of the feature itself. Red indicates high and blue indicates low. By checking the sorting of features, the distribution of shape values and color patterns, we can systematically identify the key features closely related to the potential of data assets and diagnose their contribution behavior.

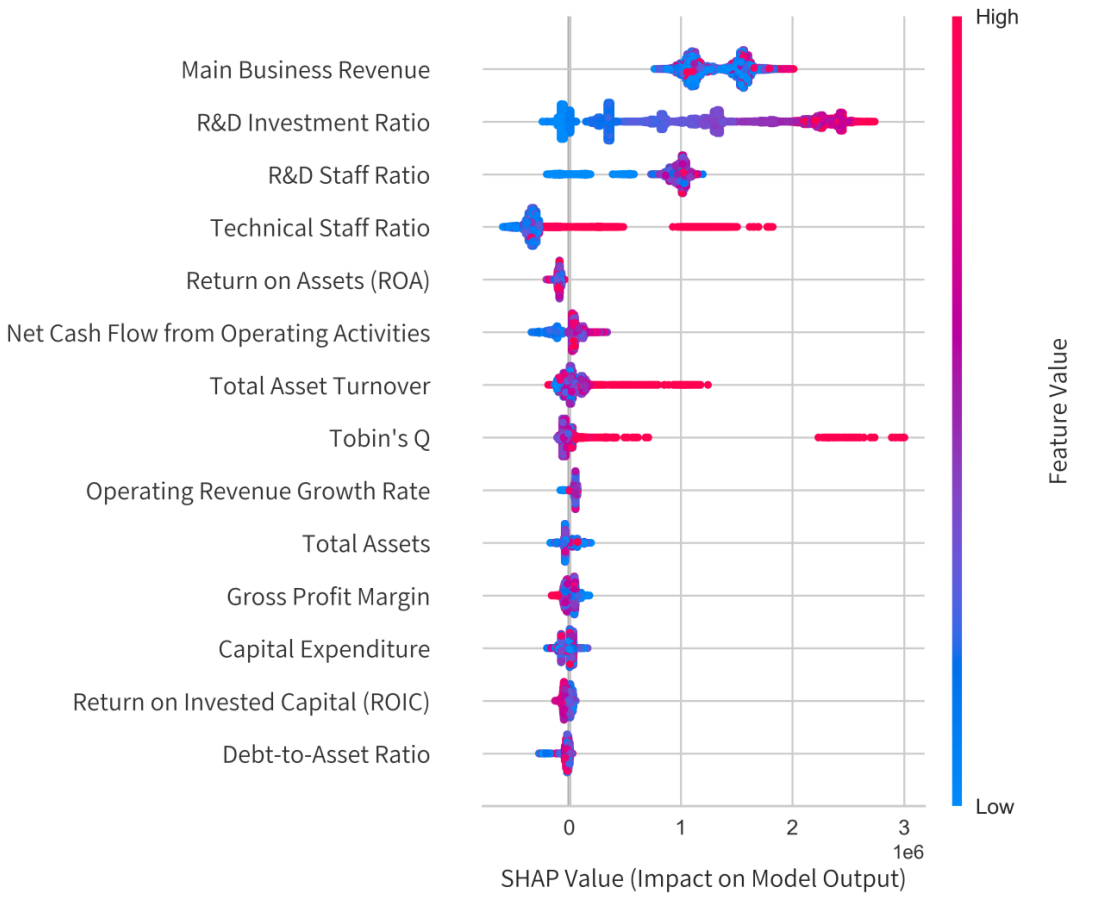

Figure 2. Contribution Analysis of Firm Characteristics to Model Output (SHAP Summary Plot)

According to the global importance ranking shown in Figure 2, the characteristics most closely related to the potential of data assets are mainly divided into two categories: innovation investment and business scale. More specifically, the operating revenue, R&D personnel ratio and R&D investment ratio rank among the top three in terms of average absolute shape value, indicating that the model heavily relies on these characteristics when making predictions. From the perspective of contribution mode, high operating income (red dots) tend to be located in the positive contribution area, which means that the larger business scale is usually related to the higher forecast value. Similarly, the high values of R&D personnel ratio and R&D investment ratio are also concentrated in the positive region, indicating that R&D intensity has a stable positive marginal effect on prediction after controlling other factors.

In addition to these core characteristics, the ratio of technical personnel, net operating cash flow and total assets are also high in the global importance ranking, and their shap value is mainly positive. This shows that sufficient liquidity and large asset scale are the supplementary characteristics of enterprises with high data asset potential. It is worth noting that the contribution mode of return on assets (ROA) is asymmetric: low ROA value (blue dot) is closely related to large negative shap value, while high ROA value (red dot) only produces relatively moderate positive shap value. This means that in the reasoning

of the model, weak profitability is a clear negative signal, while strong profitability is not the core positive signal of data asset potential.

In contrast, the global importance of financial indicators such as return on investment capital, capital expenditure and asset liability ratio is low, and their shap values are concentrated around zero. This shows that once the characteristics of innovation and scale are considered, these financial variables add only limited additional information. To sum up, the shap summary chart reveals a multi-level structure of correlation characteristics: the positive characteristics of prominent innovation investment as the core; Business scale and operational stability constitute an important foundation layer; Profitability plays an asymmetric regulatory role. The next step of the analysis will further examine the nonlinear patterns of these key features using the shape correlation graph.

3.2 Non-Linear Patterns and Interaction Effects of Associated Characteristics

In order to further explore the nonlinear relationship and interaction effect between key enterprise characteristics and data asset potential, we turn to shape dependence graph. The results are shown in Figure 3. The correlation diagram shows how the value of a single feature affects its marginal contribution to the model prediction and is displayed on a two-dimensional plane. Specifically, the x-axis displays the normalized value of the feature, and the y-axis gives the corresponding shap value, that is, the marginal contribution of the feature to the prediction at this value. Each point represents a fixed sample, and its color represents the variable value with the strongest interaction with the focus feature (blue indicates low, red indicates high). By examining the scattering patterns and color gradients, we can accurately describe how the marginal contribution of each feature varies within its range.

The results show that the characteristics related to innovation investment have obvious nonlinear threshold effect. For the R&D investment ratio, R&D personnel ratio and technical personnel ratio, when the eigenvalue is low, the shap value remains low. However, once the function exceeds a certain threshold, the shap value will rise sharply to a higher positive platform, and then tend to be stable. This model reflects the basic logic of the model: sporadic and low-intensity innovation investment is not a powerful indicator of potential, while sustained and high-intensity strategic investment is essential to realize the potential of high data assets. Interaction analysis further improves this view. In the dependence chart of ROA and return on investment (ROIC), the samples with low R&D investment (blue) tend to gather in the area where the value of shap increases with the increase of profitability. In contrast, the samples with high R&D investment (red) are more scattered in regions with high profitability but relatively low shap value. This shows that for companies that invest heavily in R&D, the current profitability is not a reliable positive signal of the potential of data assets.

Business scale and operating efficiency indicators such as operating income, total assets and total asset turnover also show a non-linear pattern. For operating income and total assets, the shap value gradually increases in the middle and low range, and then stabilizes at a higher value, indicating that the marginal contribution decreases. For the total asset turnover rate, the shap value only enters the positive contribution area, and is stable after the characteristics exceed a certain level. For some traditional

financial indicators, including gross profit margin, debt asset ratio and capital expenditure, the dependence chart shows a decentralized and multi cluster structure, without a stable monotonous trend. This means that the marginal contribution of these variables largely depends on the values of other characteristics, which mainly exert influence through complex interactions.

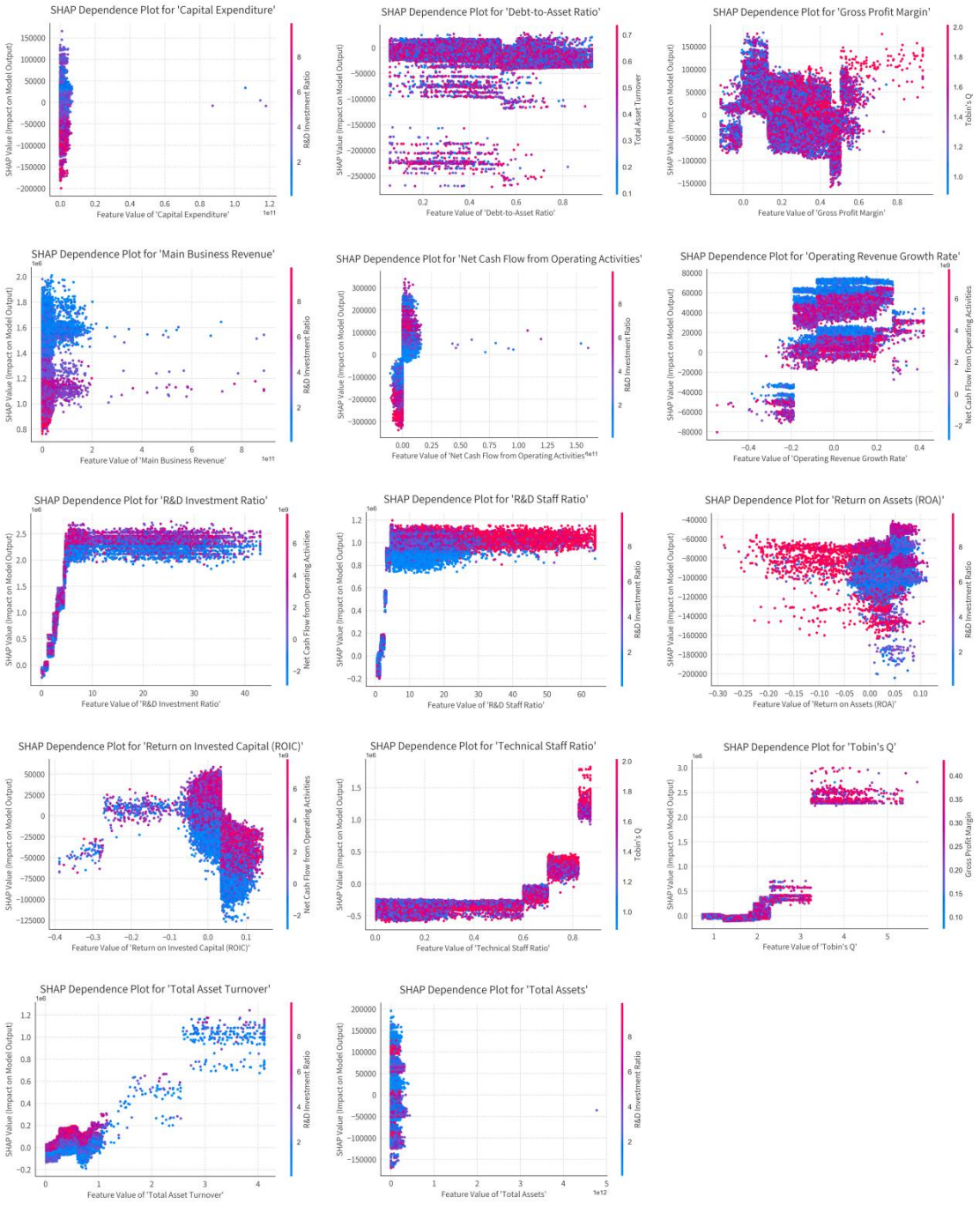

Figure 3. SHAP Dependence Plots

In short, the shape dependence graph reveals a complex set of nonlinear patterns that link enterprise characteristics with data asset potential. Innovation investment shows obvious threshold effect, and the marginal contribution of enterprise scale decreases, and the contribution of some financial indicators largely depends on the level of other variables. These findings show that the potential of data assets is not generated by the simple linear accumulation of single factors. On the contrary, it reflects the compound results formed by multidimensional interactions and characteristics, which operate differently within a specific value range.

3.3 Market-Wide Data Asset Potential Scan

After identifying the micro level characteristics related to the potential of data assets through the shap framework, we now apply the same identification framework to shift the analysis from a single company to the industry level. Using the trained and finalized lightgbm model, we scanned the potential of all A-share listed companies in order to explore how the potential of data assets is distributed across industries.

(1) Data Sources, Processing, and Descriptive Statistics

The data used for the market-wide prediction come from the CSMAR database for the year 2024. The initial sample was processed according to the following steps: all ST and *ST firms were excluded; observations with missing values on key variables were removed; and only companies belonging to the 13 industry categories that match the seed sample were kept. To reduce the influence of extreme outliers, all continuous variables were winsorized at the 1st and 99th percentiles.

After these processing steps, the final sample contains 4,460 firm-year observations. Descriptive statistics for the main variables are presented in Table 6.

Table 6. Descriptive Statistics of Market-Wide Sample

Variable Symbol	N	Mean	Sd	Min	P50	Max
Size	4460	12,960,000,000	31,640,000,000	395,600,000	3,578,000,000	2.259E+11
Revenue	4460	7,539,000,000	19,760,000,000	124,300,000	1,706,000,000	1.441E+11
Turnover	4460	0.563	0.357	0.079	0.489	2.203
Capex	4460	565,900,000	1,559,000,000	984,993.7	128,300,000	11,740,000,000
OCF	4460	697,700,000	2,109,000,000	-978,700,000	132,700,000	15,900,000,000
Growth	4460	0.046	0.239	-0.547	0.033	0.977
ROA	4460	0.018	0.067	-0.246	0.026	0.161

Variable Symbol	N	Mean	Sd	Min	P50	Max
ROIC	4460	0.023	0.085	-0.361	0.035	0.205
GPM	4460	0.280	0.182	-0.067	0.246	0.846
TobinQ	4460	2.037	1.172	0.849	1.661	7.571
TechRatio	4460	0.249	0.194	0.017	0.182	0.889
RDRatio	4460	7.343	8.306	0.060	4.750	49.970
RDPsnRatio	4460	18.429	13.689	0.510	14.695	72.730
Leverage	4460	0.397	0.204	0.055	0.385	0.915

(2) Market-Wide Potential Scan and Industry Distribution

In order to test the distribution of data asset potential in different industries, we applied the finally determined lightgbm model to predict the potential value of 4460 listed companies. The forecast values are then aggregated by industry to obtain the industry average. Figure 4 ranks the industry according to the average forecast data asset value.

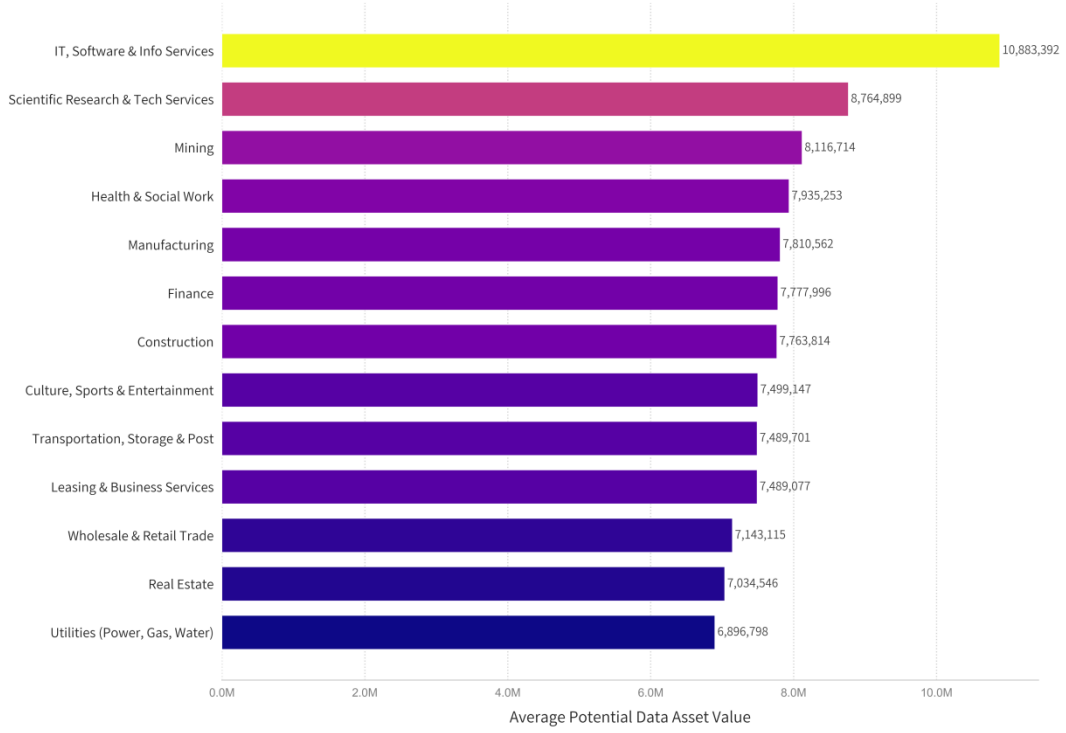

Figure 4. Ranking of Predicted Average Data Asset Values by Industry for A-Share Companies Based on LightGBM Model

The results show that the data asset potential of A-share listed companies has significant industry heterogeneity, which is closely related to the differences in technology strength, data dependence and business models among industries. Specifically, the information transmission, software and information technology services industries rank first, and their average potential value far exceeds that of all other industries. Followed by scientific research and technical services, ranking second. This distribution at the macro level is consistent with the results of the shap study at the micro level, because these two departments have the highest concentration of high-intensity innovation investment and technical talents - the core positive signal of model identification. This consistency shows that our identification framework successfully captures key industry level commonalities related to the potential of data assets. The health and social work industries also rank relatively high. This may be because it has a large number of valuable medical and health data, and the proportion of professional and technical personnel is very high, both of which are in line with the positive signals determined by the model. Manufacturing, as a broad category, is in the middle of the ranking, but this may hide considerable internal changes. For example, technology intensive sub sectors such as advanced equipment manufacturing and biopharmaceuticals may have higher data asset potential than traditional labor-intensive manufacturing industries. Due to the limited size of seed samples, this study can not further subdivide the manufacturing industry into more detailed sub sectors.

It is worth noting that capital intensive mining industry ranks third. This result may reflect the model's emphasis on business scale and the potential value of industrial data assets in mining operations. However, it may also be affected by data limitations: there is only one mining company in the seed sample, and the model may have summarized the unique characteristics of a single company, leading to possible prediction bias. At the same time, the ranking of industries such as finance and construction reflects their scale effect and the extent to which they adopt data-driven practices.

In short, the scanning of the market scope verifies the micro level correlation learned by the identification model, and confirms that the potential of data assets is mainly concentrated in innovation intensive and technology intensive industries. At the same time, it also reveals the challenge of applying the model to traditional industries, and emphasizes the potential deviation caused by the seed sample limit. These findings emphasize the importance of careful interpretation when using machine learning models trained with small samples.

3.4 Main Findings

The main findings of this study are summarized as follows.

According to the model, innovation investment and business scale are the core characteristics related to the potential of data assets.

Shap analysis shows that the innovation related indicators, especially the R&D investment rate and the proportion of R&D personnel, are the characteristics that contribute the most to the model prediction. Business scale (obtained from operating income and total assets) and operational stability (reflected in net operating cash flow) constitute a set of basic characteristics, which are closely related to high forecast

value. In contrast, traditional financial indicators such as profitability show asymmetric effects: low profitability significantly contributes negatively, while high profitability only provides moderate positive contributions.

The correlation pattern between key features and data asset potential shows significant threshold effect and interaction pattern.

The study found that the shape contribution of innovation investment follows a clear threshold model: only after reaching a critical level, the positive impact of innovation investment on the prediction will rise sharply. In addition, the model reveals complex interactions. For example, the direction and size of profitability contribution largely depend on the level of R&D investment. In enterprises with high R&D expenditure, the current profitability is even negatively correlated with the predicted data asset potential. This contrasts with the traditional valuation thinking, and emphasizes the need for a forward-looking perspective when evaluating the off balance sheet data assets of innovation driven companies.

The potential of data assets is mainly concentrated in knowledge intensive and technology intensive industries.

The scanning of market scope reveals the obvious concentration mode of specific industries. The average forecast value of information transmission, software and information technology services, as well as scientific research and technology services is much higher than that of other industries. This is consistent with the model that identifies innovation investment as the core feature.

4. Discussion and Implications

4.1 Theoretical Implications

(1) The analysis framework combining data expansion and interpretable machine learning helps to alleviate the over fitting and related problems in small sample training

Training complex machine learning models on small samples, such as lightgbm, often leads to over fitting, unstable results, unreliable estimation of feature importance, poor generalization ability and high measurement variance. In order to overcome these difficulties, this study constructs an analysis framework integrating data expansion and interpretable machine learning, and uses this framework to systematically identify the key features related to the potential of data assets. Under this framework, this study not only takes innovation investment and business scale as the core related factors of data asset potential, but also reveals the non-linearity and interaction mode between them through shap analysis. In addition, this paper also reveals the significant industry heterogeneity of data asset potential, which provides new empirical evidence for how to reflect the value of data assets.

(2) The analytical framework developed here can also be extended to scan key features across the full sample population

By applying the trained and finalized LightGBM model together with the SHAP identification framework, the analysis moves from the level of individual firms to that of industries, conducting a market-wide scan of data asset potential among A-share listed companies. The results show clear industry heterogeneity in

data asset potential, and its distribution is closely linked to each industry's technological intensity, reliance on data, and business model characteristics.

4.2 Practical Implications

The analytical framework and the empirical findings concerning the associated characteristics and industry heterogeneity of data asset potential have both theoretical and practical value. They provide useful decision-making references and action guidelines for various stakeholders in the data element market.

(1) Impact on the development of data element market system

First, regulators should encourage enterprises to disclose key non-financial information to improve the transparency of off balance sheet data assets. Our research finds that innovation investment and business scale are closely related to the potential of data assets, and there is significant industry heterogeneity. This provides a basis for regulators to encourage and guide enterprises to disclose more non-financial information about R&D expenditure, talent structure and data governance. This will enhance the ability of the market to identify off balance sheet data assets and reduce information asymmetry.

Second, the government should tilt its industrial policies to encourage long-term strategic investment. The threshold effect of innovation investment means that continuous high-intensity investment is crucial to the construction of high-value data assets. Therefore, industrial policies aimed at supporting digital transformation should shift from broad-based support to encouraging long-term strategic investment, and turn social capital and enterprise resources to activities that generate core data assets.

(2) Impact on asset appraisal practice

First of all, asset appraisal institutions and appraisers should actively meet the opportunities and challenges brought by data asset appraisal. The capitalization of data assets is a new knowledge intensive field of valuation, which poses a challenge to the existing knowledge base and method system. Evaluation institutions should take data asset evaluation as a strategic service line and organize systematic training on data science, industrial economics and enterprise innovation strategy. Appraisers need to update their knowledge to understand the unique role of data elements in value creation, and meet the resulting professional challenges, so that they can make a fair and reliable assessment.

Second, asset evaluation can adopt the situational analysis perspective to realize the comprehensive evaluation of data assets. When evaluating the contribution of traditional financial indicators such as profitability to the value of data assets, appraisers must consider the company's R&D investment level and its industry background. The type and stock of data assets vary from industry to industry, which requires a case study based on the business model and size of each company. This ability will become the core competence of future data asset appraisers.

(3) Enlightenment for investors' decision making and value discovery

First, investors should establish a more effective value discovery framework to improve the efficiency of their investment decisions. In the digital economy, enterprise value exceeds current profits. The indicators determined in this study, such as R&D investment ratio and labor structure, can be used as effective

signals to identify high potential enterprises, so as to make a more forward-looking judgment on long-term value.

Second, investors should carefully evaluate the strategic investment in innovative enterprises. This study provides evidence to help solve the obvious tension between high R&D expenditure and low liquidity profits. This indicates that strategic investment to obtain long-term data advantages may temporarily reduce short-term profits. Investors should avoid missing high-quality investment opportunities just because they are worried about short-term profitability.

(4) Impact on enterprise data strategy and internal governance

First, enterprises should deeply integrate their data strategy with their core innovation and business strategy. The results show that data strategy must be closely combined with innovation and business strategy. Managers should regard the cultivation of data assets as a long-term and systematic project, which requires continuous and high-intensity investment. This means that strategic resources must be allocated to R&D, talent and digital projects in the annual budget and long-term planning.

Second, enterprises should strengthen the whole chain governance and value management from data resources to data assets. Because the value of data assets is closely related to the generation and use of data, managers need to establish a comprehensive data governance system to track data sources, quality, costs, application scenarios and benefits. This is not only the basis for future capitalization and value evaluation of data assets, but also the internal requirement for transforming original data resources into real data assets.

References

Chen, Z. B., & Yang, J. (2025). Research on the Corporate Value Effects of Data Assetization—Empirical Evidence from the Promulgation of New Data Asset Capitalization Regulations. *Accounting Research*, *5*(2025), 18-30.

He, Y., Chen, L. L., & Du, Y. G. (2024). Can Data Assetization Alleviate Financing Constraints for "Specialized, Refined, Distinctive, and Innovative" SMEs. *China Industrial Economics*, *8*(2024), 154-173.

He, Y., Ren, L. Q., Yu, W. L., & Du Y. G. (2024). Firm and Executive Characteristics and Corporate Misconduct—Empirical Evidence Based on Machine Learning. *Journal of Management Sciences in China*, *27*,6(2024), 43-68.

Huang, C., Shi, Z. Q., & Jiang, Q. S. (2024). The Impact of Data Asset Capitalization on Firms: Evidence from the Stock Market. *Management Science*, *37*,6(2024), 43-61.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*, 5(2020), 2223-2273. https://doi.org/10.1093/rfs/hhaa009

Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, *110*, 9(2020), 2819-2858. https://doi.org/10.1257/aer.20191330

Ke, G. et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*(2017), 209-227.

Li, J. W., Hu, W. Y., & Wang, W. (2025). Factors Influencing Corporate Digital Transformation and Their Contributions—Research Based on the LightGBM Algorithm. *Science and Technology Management Research*, *45*,09(2025), 64-73.

Li, J., Dong, X. F., Zhang, J. L., & Tao, Y. Q. (2023). The Impact of Data Assets on Corporate Innovation Input. Foreign Economics & Management, *45*,12(2023), 18-33.

Li, S. G., Shao, H. B., Fang, F., & Lu, F. C. (2025). Corporate Data Asset Information Disclosure and Capital Market Pricing Efficiency. *China Industrial Economics*, *7*(2025), 138-155.

Lu, Y. et al. (2020). Executive Characteristics and Firm Performance—Empirical Evidence Based on Machine Learning. *Journal of Management Sciences in China*, *23*, 2(2020), 120-140.

Luo, M., Li, J. P., & Tang, K. (2023). Corporate Data Assetization: Accounting Recognition and Valuation. *Journal of Tsinghua University* (Philosophy and Social Sciences), *38*, 5(2023): 195-209+226.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*(2017), 309-336.

Lundberg, S. M. et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, *2*(2020), 56-67. https://doi.org/10.1038/s42256-019-0138-9

Mao, C. M. et al. (2025). How Data Assets Enhance Corporate Green Innovation Capabilities—Based on Causal Inference Using Double Machine Learning Models[J/OL]. *Science and Technology Progress and Policy*, 8(2025), 1-10.

Maximilian, S., Huang, Y. L., Kupfer, A., & Krcmar, H. (2024). The Effect of Digital Platform Strategies on Firm Value in the Banking Industry. *Journal of Management Information Systems*, *41*(2024), 2, 394-421. https://doi.org/10.1080/07421222.2024.2340825

Mobina, Z., Najah, A., Sadok El, G., Iraj, F. (2024). Firm digital transformation and corporate performance: The moderating effect of organizational capital. *Finance Research Letters*, *61*(2024), 1-10. https://doi.org/10.1016/j.frl.2024.105032

Niu, B., Du, Y. Q., Yu, X., & Zhao, N. H. (2024). Data Asset Information Disclosure and Bond Financing Costs. Journal of Guangdong University of Finance and Economics, 39, 05(2024), 88-101.

Ren, M. D., Yang, L., Wu, Y. R. (2025). Analysis of Factors and Implementation Paths for Data Asset Capitalization—Based on a Multi-Case Study. *Accounting Newsletter*, *7*(2025), 90-100.

Xu, L. et al. (2019). Modeling tabular data using conditional gan. *Advances in neural information processing systems*, *32*(2019), 67-81.

Xu, X. C., Zhang, Z. W., & Hu, Y. R. (2022). Research on Data Asset Statistics and Accounting. Management World, 38,02(2022), 16-30+2.

Yoshioka, T. (2024). Valuation of Intangible Fixed Assets Using Generative Artificial Intelligence and Machine Learning. *Journal of Management Science*, *13*(2024), 27-36.

Zhang, J. R., Zhao, W. N., Wang, Q. W. (2024). Current Status and Market Reactions to Data Asset Capitalization by Listed Companies—Evidence from A-Share Interim Reports. *Accounting Monthly*, *45*, 24(2024), 42-50.

Zhou, Y. R., Zhang, X. Y., Liao, S. Y., & Tan, L. (2025). Analysis of Data Resource Capitalization by Companies Listed on Data Exchanges—Taking COSCO Shipping Technology as an Example. *Accounting Monthly*, *46*,14(2025), 104-109.