Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on the Construction of Enterprise Data Asset Pricing Model Based on Machine Learning in Digital Economy Environment

Yongxia Lv[1,] ✉

[1] *Faculty of Accounting, Zhengzhou Vocational College of Finance and Taxation, Zhengzhou, Henan, 450048, China*

ABSTRACT

In order to solve the enterprise data asset pricing problem in the digital economy environment, this paper utilizes machine learning algorithms such as multiple regression model, BP neural network, and random forest regression, respectively, to price enterprise data assets. Subsequently, the data obtained from each model is fused using the integrated Stacking algorithm to construct an enterprise data asset pricing model with integrated machine learning algorithms. Predictive estimation of the pricing of enterprise data assets is carried out after a detailed justification of the parameter selection of the model. The results show that data capacity, size, quality and freshness are the main influences on data asset pricing. The results of the parameter investigation show that the overall performance of the model is best when the number of node features is 7, at which time the explanatory degree and goodness of fit of the model are 94.33% and 97.27%, respectively. The accuracy, precision, recall and F1 value of the Stacking-based fusion model for enterprise data asset pricing prediction model increased by about 10% compared to the other three models, respectively, to achieve accurate pricing of enterprise data assets.

Keywords: Data asset pricing, random forest analysis, multiple linear regression, BP neural network model, Stacking fusion model

1. Introduction

In the current context of digital economy, data, as the smallest unit of information, once became an important resource for economic development by virtue of its fundamental, strategic and critical characteristics [1-2]. In order to promote the vigorous development of the data factor market, the state has issued a series of policy documents, aiming to promote the rational allocation and efficient utilization of data resources through the joint role of policy guidance and market mechanism [3-4]. The strategic significance of big data is no longer limited to individual enterprises or specific industries,

✉ Corresponding author.
E-mail address: lyx507hu@163.com (Y. Lv).
Received 03 March 2024; Revised 12 April 2024; Accepted 15 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-234

but has been elevated to the national level [5]. Numerous enterprises are actively responding to this trend by digging deeper and utilizing their data assets through informatization and digital transformation to achieve greater economic value [6-7]. However, data assets are significantly different from traditional assets due to their characteristics, and it is difficult to form a standardized and unified method to accurately and valuation and pricing of them. This situation brings challenges to the accounting recognition and statement presentation of data assets. Therefore, it is especially critical to accurately assess the value of data assets [8-10].

With the continuous progress of big data and artificial intelligence technology, all industries are experiencing the process of digital transformation, and the importance of reasonable valuation of data assets is becoming more and more prominent [11-12]. On the one hand, scientific valuation of data assets helps to improve the data management efficiency of enterprises. The massive data resources generated in the daily operation of enterprises can be transformed into data assets with asset attributes after appropriate processing [13-14]. Reasonable valuation of these assets can enhance the attention of enterprise managers to data assets, which in turn inspires enterprises to deeply analyze the data and use technical means to transform the data into decision support, thus creating economic value for the enterprise [15-16]. Through scientific management and utilization of data assets, not only can they maintain their value, but are even expected to realize their value-added, thus enhancing the rationality of enterprise decision-making and economic benefits. On the other hand, the construction of a perfect data asset value assessment model has certain significance for the steady and healthy development of enterprises [17-18]. Taking enterprise mergers and acquisitions as an example, the model can more accurately reflect the actual value of the enterprise and provide a strong basis for the determination of the transaction price [19-20]. At the same time, the model also helps investors to assess the overall and dynamic value of enterprise assets more scientifically and comprehensively, so as to more accurately reflect the real situation of enterprises. This will help enterprises obtain a more fair evaluation in the capital market and lay a solid foundation for their long-term development [21-22].

The current research on data asset connotation analysis and pricing is mainly combing and summarizing research, which mainly optimizes the existing data asset pricing models and methods through the connotation of data asset value and data asset features. Li, Y analyzed and combed the definition of data assets and accounting treatment of data assets as well as the pricing problem, which mainly includes mining the features of data assets and incorporating them into the accounting treatment and asset quantification framework, the study enriches the pricing theory of digital assets [23]. Yanlin, W et al. attempted to analyze the connotation and basic characteristics of data assets from the perspective of digital asset value, summarized relevant studies on digital asset value assessment, and looked forward to the future research direction of data asset pricing by focusing on natural value-addedness, externality, and multi-dimensional value characteristics of digital assets [24]. Badewitz, W et al. affirmed the value of enterprise digital assets and explored the issues related to enterprise digital asset pricing, and synthesized the research related to enterprise digital asset pricing, which laid the foundation for future research on enterprise data asset pricing [25]. There are also a considerable number of researchers who have conducted research around the sharing of enterprise data assets, integration and empowerment issues, digital asset disclosure issues, and data asset quality risk assessment models, which can help to further understand the value and connotation of data assets, and then provide an important reference for data assets. Fernandez, R. C, et al. illustrated that data can be combined with specialized knowledge and resources to generate value but It is difficult for enterprises to share, discover, and integrate data assets, so it is proposed to use a data marketplace platform to solve the related problems, and it is also pointed out that it is very important to expand the data to each organization as an asset [26]. You, J et al. envisioned a data asset quality risk assessment model with improved FMEA as the core logic, and introduced the assessment model into real cases for analysis and demonstration, confirming the effectiveness of the proposed model [27]. Li,

Y et al. explored the issues related to the disclosure of corporate data assets and concluded that the disclosure of corporate data assets can be used as a basis for non-professional investors to evaluate and judge the non-financial performance and investment value of the enterprise, which can help to enhance the value of corporate investment, and also, from the perspective of regulations, elucidated that the mandatory disclosure of digital assets can help to improve the problem of asymmetry of information in the capital market, and to protect the interests of non-professional investors [28].

This paper selects the pricing problem of enterprise data asset value as the entry point, takes the enterprise data asset as the research object, and constructs three basic machine learning prediction models, namely, "multiple linear regression model, BP neural network model, and random forest regression algorithm model" for pricing the enterprise data asset. After that, we use the integrated Stacking fusion model to construct a higher-level combination model on top of the three basic models. By comparing the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Goodness of Fit (R^2) of the prediction results of the four models, a suitable machine learning model for corporate asset pricing is selected.

2. Machine learning based enterprise data asset pricing model construction

2.1. Multiple linear regression modeling

When fitting a multivariate linear regression model [29] (LM), it is first necessary to determine the sample characteristics and prediction objectives, in this paper, the data capacity, data size, freshness, data quality, data volume score of five influencing factors as the model of the sample characteristics of the model, enterprise data asset pricing for the prediction objectives, the specific form of the model is shown below:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_0 \tag{1}$$

where y is the data asset pricing, $x_1, x_2, \ldots\ldots x_5$ is the data capacity, data size, freshness, data quality, and data volume scores, respectively, β_0 is the error term, and $\beta_1, \beta_2, \ldots\ldots \beta_5$ is the regression coefficient of the corresponding sample characteristics, respectively. The most important thing to fit the multiple linear regression model is to obtain the regression coefficients and error terms, in this paper, we use the least squares method to obtain the regression coefficients and error terms of the multiple linear regression model, and the goal of the least squares method is to minimize the residual sum of squares, and its solution formula is:

$$(X'X)\beta = X'y \tag{2}$$

where X is a matrix with m rows and n columns, m is the number of training sets, n is the number of sample features, where the first column is all 1s, indicating the intercept; X' is the transpose matrix of X; y is a vector of dependent variables with m rows and 1 column; β is a vector of coefficients with n rows and 1 column.

In this paper, we use Python to write a least squares method to construct a multiple linear regression data asset pricing prediction model for each firm by substituting the training set of 10 firms in Province A. The regression coefficients and error terms of the model are expressed in the form of regression equations.

2.2. BP neural network based prediction models

2.2.1. Neuron Modeling. A typical neuronal model consists of five components:

1) An n-dimensional input variable;

2) An output variable;

3) A set of connection weights corresponding to the input variable, reflecting the weight of the input

variable;
4) a unit of summation, a weighted sum of the input variables with their corresponding weights, and finally a threshold value;
5) An activation function, usually a nonlinear function, whose main role is to realize the nonlinear transformation of data and solve the problem of insufficient capacity of linear models.

The neural network metamodel output values are expressed as follows:

$$y = f\left(\sum_{i=1}^{n+1} x_i * w_i\right) = f\left(\sum_{i=1}^{n} x_i * w_i - \theta\right) \tag{3}$$

Where: x_i : denotes the i nd dimensional input variable.
w_i : denotes the weight corresponding to the i th dimensional input variable;
θ : denotes the weights of this neuron;
y : denotes the output value of this neural unit model;
$f(\cdot)$: denotes the activation function.

2.2.2. BP Neural Network Model Structure. The BP neural network model [30] is a multilayer feed-forward network trained according to the error back-propagation algorithm, and the model structure usually consists of three parts: an input layer, a hidden layer, and an output layer. Usually the number of nodes in the input and output layers of the network is determined, while the number of nodes in the hidden layer is uncertain, and the optimal number of nodes needs to be determined according to the specific problem. The BP neural network model structure of a single hidden layer carries a nonlinear activation function in addition to the input nodes to realize the nonlinear transformation of data.

Learning Algorithm Flow

Based on the above introduction of the neural network model structure and BP learning principle, there is a basic understanding of the neural network algorithm, the following will be a specific introduction to the neural network learning algorithm process:
1) Randomly initialize the weight coefficient $W = (W_1, W_2, \ldots\ldots W_n, W_{n+1})$ from the input layer to the output layer, usually a smaller zero random number, denoted as $W_1(0), W_2(0) \ldots\ldots W_n(0)$. Also $W_{n+1} = -\theta$. Where $W_i(t)$ is the threshold at moment t .
2) Input sample $X = (X_1, X_2, \ldots\ldots X_n, X_{n+1})$ and expect output D ;
3) Calculate the actual output according to the neuron model formula:

$$y_i = f\left(\sum_{i=1}^{n+1} X_i * W_i(t)\right) \tag{4}$$

4) Find the error function e from the actual output:

$$e = D - Y(t) \tag{5}$$

5) Modify the weights with error function e :

$$W_i(t+1) = W_i(t) + \eta * e * X_i \tag{6}$$

Where, $i = 1, 2, \ldots\ldots n, n+1$, η is the learning factor, which is calculated by various types of learning algorithms, and cannot be too large, which may make the error function not converge during the training process, or too small, which may lead to too slow training speed. When the actual output and the desired output are the same there:

$$W_i(t+1) = W_i(t) \tag{7}$$

6) Repeat the above steps until the training results are stabilized and output.

The flow of neural network BP learning algorithm is shown in Fig. 1.

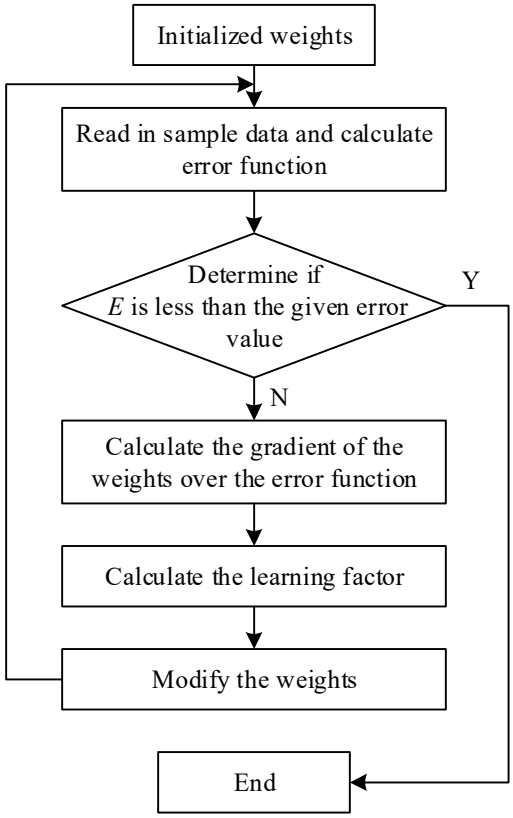


Fig. 1. Neural network bp learning algorithm process

2.2.3. Predictive modeling:

1) Determination of model variables. According to the establishment of BP neural network prediction model, "data capacity, data size, freshness, data quality, data volume score" of enterprise data assets at the current moment is defined as the input variable (independent variable). The enterprise data asset pricing is the output variable (dependent variable). The quantitative description of the model's influencing factors is shown in Table 1.

Table 1. Model impact factor quantification

Symbol	Influencing factor
X1	Data capacity
X2	Data size
X3	freshness
X4	Data quality
X5	Data score

2) BP neural network model design. The realization of the BP neural network model prediction requires the design and determination of its components and related parameters, and due to the complexity of the model structure, the design will be discussed in detail below.

The number of model layers and nodes are determined:

The hidden layer can have one or more layers, considering the principle of using the simplest model to achieve the best results, the number of hidden layers is determined to be 1 layer.

The number of nodes in the hidden layer will have a large impact on the performance of the neural network, and the more the number of nodes, the higher the training accuracy of the model will be. For the determination of the number of nodes in the hidden layer, most of the current determination of the number of nodes in the hidden layer is based on empirical formulas, such as: $2n+1$ method,

$\sqrt{n+m}+\alpha$ method, etc., in which n is the number of nodes in the input layer, m is the number of nodes in the output layer, $\alpha \in [1,10]$. However, the empirical method can't be directly applied to promote and can only determine the number of nodes in the hidden layer is inevitably related to the complexity of the intrinsic laws of the training samples, so this paper utilizes the sample data to train and observe model Performance. When the number of hidden layer nodes is 7, the model accuracy is the highest, so the number of hidden layer nodes is finally determined as 7. In this paper, we choose the current common Sigmoid activation function for testing.

Sigmoid: A type S activation function that maps the input to a value between 0 and 1, with the following functional expression:

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{8}$$

3) BP learning algorithm determination. The BP neural network utilizes the gradient search technique with a view to minimizing the mean squared error between the actual and desired output values of the network. The gradient descent method is subdivided into: batch gradient descent (GD), stochastic gradient descent (SGD) and small batch gradient descent (MBGD). The details of the neural network model are shown in Table 2.

Table 2. Details of the neural network model

Model details	Bp neural network
The number of input layer variables	5
Hidden layer activation function	RELU
Output layer activation function	Identity
Learning algorithm	Small batch gradient descent
Error function	Error sum of squares
The number of hidden layers of neurons	7
The number of output layers	1

2.3. Predictive model based on random forest regression algorithm

Random Forest [31] (RF) is an integrated learning algorithm based on multiple decision trees and is one of the most representative algorithms in Bagging algorithms.

Based on the research content of this paper, it will mainly introduce the principle of random forest regression algorithm, namely: the influence of independent variables on the dependent variable, assuming that the dependent variable is Y there are n observations, and there are M independent variables that have an influence on the dependent variable, the specific algorithmic process:

1) From the original training set of n cases, the Bootstrap method is applied to repeat the sampling of N training samples, a total of T times sampling, to generate T training sets, respectively, training T. And the T samples consisting of cases that were not sampled each time the training samples were taken were out-of-bag (OOB) data, which were used as the test sample set.

2) In constructing the CART tree model, m features were randomly selected from the M independent variables as alternative branching variables at the branching nodes of each tree.

3) Each CART tree is grown continuously without branch reduction, by setting the tree value ntree of the decision tree as a termination condition for the growth of the decision tree.

4) The generated T decision trees are composed into a random forest regression model; the model estimation effect is evaluated by the accuracy of out-of-bag data prediction, i.e., the mean square error of the test set, assuming that the number of out-of-bag data samples is k:

$$MSE_{OOB} = \frac{\sum_{i=1}^{k}\left(y_i - \hat{y}_i\right)^2}{m} \tag{9}$$

$$R_{RF}^2 = 1 - \frac{MSE_{OOB}}{\hat{\sigma}_y^2} \tag{10}$$

Where, y_i: denotes the actual value of the dependent variable in OOB; $\hat{y}_l$: denotes the predicted value of the model; $\hat{\sigma}_y^2$: denotes the variance of the OOB predicted value.
5) The mean of the predicted results from the T decision trees is used as the model output.

2.4. Stacking Fusion Model

Stacking fusion modeling [32-33] is an integrated learning approach, which is based on the principle of constructing a higher-level combinatorial model based on multiple base models. Specifically, a series of independently trained base machine learning models, which cover different types of algorithms, are first applied to predict the data. Next, the predictions produced by each base model on the training set are aggregated and treated as a new set of features to be input into a higher-level model meta-model. This meta-model generates the final prediction results by learning and optimally integrating the prediction outputs of each base model. Through this cascading and fusion of results, Stacking draws on the strengths of each base model while reducing the limitations of a single model, thus improving prediction performance. The flow principle of the Stacking fusion model is shown in Figure 2:

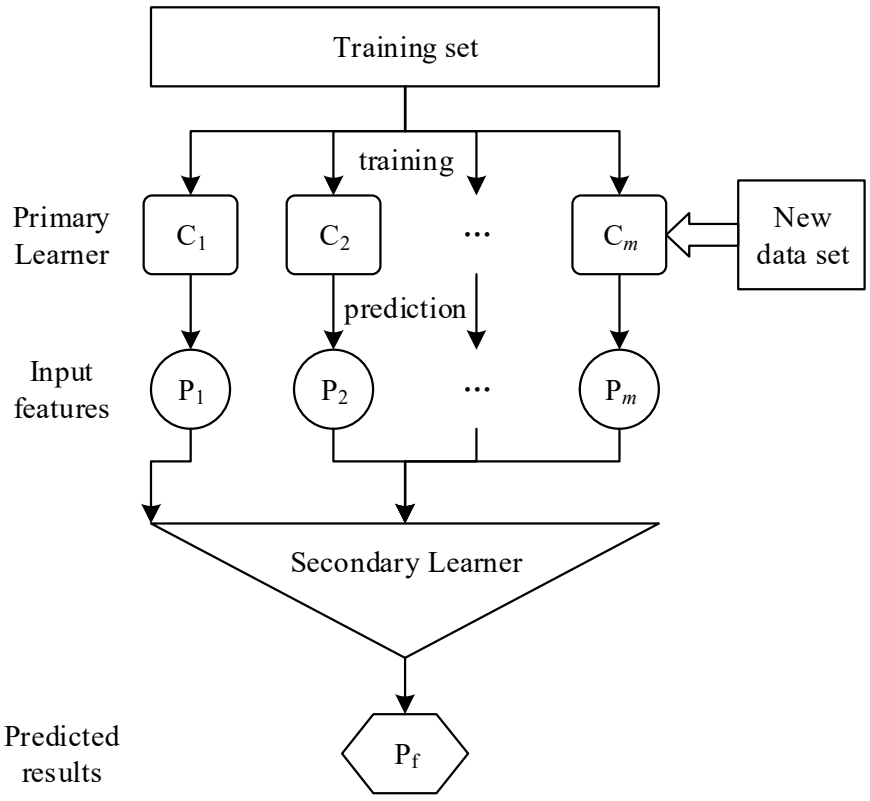


Fig. 2. Fusion model process diagram

According to the principle shown in the fusion modeling process, the above model fusion process can be explained as follows. This Stacking approach integrates the training results of each base

classifier as well as combines various relevant information that may determine the classification, which not only makes full use of the diversity of primary learners, but also further improves the performance of the overall model through the training of secondary learners, providing a powerful machine learning framework for solving complex problems.The Stacking fusion process is as follows:

Step 1: Primary Learner Training: in this phase, the training set is introduced to a diverse set of primary learners, each representing a different side of the data, labeled from C1 to C2. Such diversity ensures that a wide range of features of the data are covered.

Step 2: Cross-validation: a strong validation mechanism is provided for the training of each primary learner by employing k-fold cross-validation (k=5). Each round of cross-validation involves dividing the training set into five parts, one of which is used for validation and the remaining four for training. This approach helps the model to generalize better and reduces the risk of overfitting.

Step 3: Generating secondary learner inputs: by averaging the five predictions generated by each primary learner, the input features P1 to Pm for the secondary learners were obtained. Such averaging takes into account the differences in the different folds of cross-validation and provides more robust inputs to the secondary learners.

Step 4: Secondary Learner Training: In this step, a secondary learner is trained by using the averaged predictions of the primary learner as input features in combination with real labels. Secondary Learner Training: In this step, in this paper, a secondary learner is trained by using the average prediction result of the primary learner as an input symbol, coupled with real labels.

Step 5: Stacking: For the new dataset, the prediction results are firstly generated by the primary learner, and then these results are passed to the trained secondary learner to finally get more accurate and robust prediction results.

3. Analysis of prediction results of enterprise data asset pricing models under machine learning

3.1. Results and analysis of data processing

3.1.1. Data pre-processing. As the online data in the data trading platform will exist in the case of missing, abnormal, etc., so in the data before the application of absolute pricing method practice needs to be created for the experimental dataset of this paper to carry out pre-processing operations, for the characteristics of the experimental dataset this work carries out the following pre-processing operations. First, remove the data in the dataset API online data call times for the year, month, day data, at the same time with the total transaction price divided by the number of calls to get the API data per call unit price. Second, remove the data whose online data price is empty or 0. Third, when quantifying the data size element indicator, the data size is converted in MB24 units in a uniform manner. Fourth, the missing values present in the online data were filled, such as the median filling was used for the information entropy score, and in addition, some outliers in the dataset were deleted. Fifth, for the realization of N-way K-shot MAML absolute pricing experiments. Some information of the dataset after preprocessing is shown in Table 3. The experimental data is divided into seven categories based on platforms, experts and related literature research: industrial economy, financial services, research and technology, enterprise management, life services, marketing services and public opinion monitoring. The absolute pricing element indicators included are data size, scene coverage, time span, information entropy score, data sensitivity, data freshness, data citation and data rating.

Table 3. After pre-processed data division information

Data category	Data sensitivity	Data score	Information entropy score	Time span	Data freshness	Data reference	Data size	Data volume	price

Industrial economy	Medium	5	4.5	2	0.1003	0.4958	114.74	10036	2093.67
Financial services	Medium	5	4.5	1	0.0985	0.5013	242.11	10026	6167.52
Business management	Medium	5	4.5	1	0.0996	0.4987	65.19	10010	4390.74
Scientific research	Low	4	4	24	0.0983	0.4973	36.35	9981	8184.58
Marketing service	Medium	5	4.5	12	0.1011	0.4972	111.14	9965	8296.54
Life service	Low	4	3.5	12	0.1002	0.4994	1292.47	9995	1.1
Public opinion monitoring	Medium	5	4.5	1	0.1	0.4945	13308.59	10010	0.48

3.1.2. Absolute pricing characterization. After data preprocessing, the absolute pricing dataset contains a total of 10,218 online data asset information, which can be classified into seven categories according to different data categories: G scientific research and technology, F financial services, E industrial economy, D enterprise management, C life services, B marketing services, and A public opinion monitoring, and the statistics of the sample size of the classification of the online data assets are shown in Figure 3. Among them, the data products of scientific research and technology category have the largest number of samples, with 2,333 items, and the data products of marketing service category have the smallest number of samples, with 738 items.

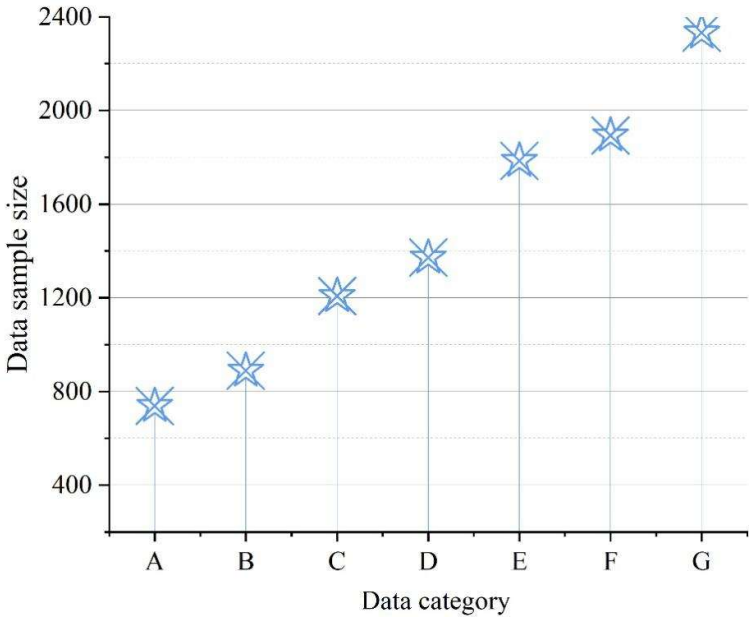


Fig. 3. The online data asset classification sample size statistics

The impact of absolute pricing element indicators on data prices does not necessarily follow an absolute linear relationship. The relationship between data score and price is shown in Figure 4. As is usually the case, data products with high data scores will have higher pricing prices, but as can be seen in the graph of the relationship between data scores and data prices shown in the figure, data products with low data scores may also have higher transaction prices. For example, the product with a data

score of 4 in the figure is trading at a price that is 0.51201 (1e6) higher than the highest price traded for a product with a score of 5.

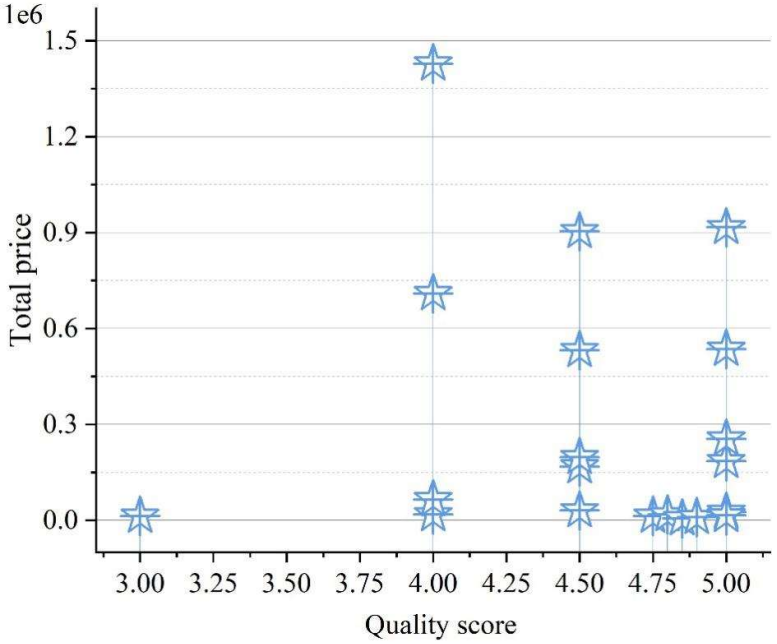


Fig. 4. Data score and price relationship

3.2. Implementation of the data asset value assessment model

3.2.1. Assessment model based on multiple linear regression. In order to compare the prediction results of machine learning with those of traditional multiple linear regression, the prediction is first modeled using multiple linear regression. In this paper, we draw on the idea of linear modeling of data assets and construct the following prediction model by combining the final selected feature variables:

$$\begin{aligned} Value = {} & \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 volume \\ & + \beta_5 size + \beta_6 freshness + \beta_1 quality \\ & + \beta_8 dvSScore + \beta_9 redundScore + \beta_{10} scareScore + \varepsilon \end{aligned} \tag{11}$$

where $Value$ is the data asset value, β_0 is the constant term, β_1, β_2 and β_3 are the regression coefficients of the dummy variables characterizing the industry classification, D_1 represents the economic and financial category, D_2 represents the scientific and technological category, D_3 represents the public opinion monitoring category, $volume$, $size$, $freshness$, $quality$, $dvScore$, $redundant\ Score$ and $scare\ Score$ are the data capacity, data size, freshness, data quality, data volume score, redundancy score and scarcity score, and the coefficients in front of each of the variables are the regression coefficients. Coefficients in front of the variables are regression coefficients, and ε is the residual term. In this paper, we name the top 7 data based on the importance of the features of Random Forest, as shown below: the names of the features ranked 1-7 are "data capacity, data size, freshness, data quality, scientific research and technology, public opinion monitoring, and economy and finance" respectively.

The results of the multiple linear regression are shown in Table 4. The regression results show that the value of data assets is significantly correlated with data capacity, data size, data quality and freshness regardless of other factors, indicating that these four features are the main influencing factors on the value of data assets. This finding is consistent with the results based on random forest feature selection, further validating the accuracy of feature selection. For the prediction accuracy of the multiple linear regression model, the goodness-of-fit R^2 of the training set is 21.32%, while the

goodness-of-fit R^2 of the test set is 19.46%, which is not satisfactory, suggesting that there may not be a linear relationship between the value of the data asset and the influencing factors. In addition, the high volatility of data asset prices and not fully conforming to normal distribution may also contribute to the poor prediction effect of the multiple linear regression model. However, the correlation does not affect the conclusion that data capacity, data size, data quality, freshness, and the industry to which it belongs are important influences on the value of data assets.

Table 4. Multivariate linear regression

	Dependent variable						
	1	2	3	4	5	6	7
D_1	0.0211***	0.0176***	0.0190***	0.0192***	0.0191***	0.0191***	0.0191***
	(4.2652)	(3.5338)	(4.0896)	(3.9386)	(3.9164)	(3.9012)	(3.9099)
D_2	0.0735***	0.0592***	0.0618***	0.0613***	0.0614***	0.0615***	0.0615***
	(14.7029)	(11.9840)	(12.3448)	(12.2668)	(12.2777)	(12.3045)	(12.3419)
D_3	0.0038	-0.0025	-0.0039	-0.0037	-0.0037	-0.0037	-0.0037
	(0.697)	(-0.4852)	(-0.7235)	(-0.6832)	(-0.6818)	(-0.6952)	(-0.6988)
volume	0.1077***	0.1013***	0.1014***	0.0995***	0.0994***	0.0996***	0.0994***
	(12.0901)	(11.631)	(11.6706)	(11.5722)	(11.5642)	(11.6109)	(11.5838)
size		0.1951***	0.1939***	0.1932***	0.1934***	0.1939***	0.1941***
		(12.3335)	(12.2736)	(12.2738)	(12.2456)	(12.3248)	(12.3301)
freshness			-0.0434***	-0.0426***	-0.0422***	-0.0403***	-0.0408***
			(-3.9279)	(-3.7721)	(-3.7331)	(-3.6222)	(-3.5984)
quality				0.0388***	0.0358***	0.0439***	0.0399***
				(4.3243)	(3.7361)	(4.2012)	(3.2989)
dvScore					0.0041	0.0048	0.0035
					(0.8178)	(0.4648)	(0.6283)
redundScore						-0.0082*	-0.0075*
						(-1.9354)	(-1.6985)
scarceScor							0.0046
							(0.8878)
Constant	0.0013	0.0013	0.0081*	-0.0113*	-0.0115*	-0.0116*	-0.0117*
	(0.2806)	(0.2842)	(1.8078)	(-1.7835)	(-1.8093)	(-1.8396)	(-1.8413)
Observations	2803	2803	2803	2803	2803	2803	2803
Adjusted R^2	0.1687	0.2125	0.2167	0.2218	0.2217	0.2225	0.2224

3.2.2. Evaluation model based on BP neural network. In order to determine the appropriate parameters, 1~3 hidden layers were set sequentially, the corresponding number of nodes were trained from 5~6, the learning rate was set to 0.0005 with an initial threshold of 0.05, the model was trained

using the training set, the model accuracy was verified with the test set, and the optimal model was selected by calculating and comparing the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the Goodness-of-Fit (R^2),. The neural network model training results are shown in Table 5. From the table, it can be seen that the prediction accuracy of the double hidden layer model is better than the single hidden layer model. In terms of Root Mean Square Error (RMSE), the double hidden layer model with the number of nodes (5,4) is the lowest at 681.4972 and is significantly lower than the other models while its goodness of fit (R^2) is also the highest among all the models at 94.82%. In terms of Mean Absolute Error (MAE), the single hidden layer model with the number of nodes 3 is the lowest at 243.4921 and its advantage over the double hidden layer model with the number of nodes (5,4) at 248.5108 is not significant. Finally, in terms of running time, the double hidden layer model with (5,4) nodes has a training time of only 12.29 minutes, which is not the longest despite the fact that the model itself is more complex than the single layer model and has the highest number of nodes. The focus of this paper is on the prediction accuracy of the model, so in summary, the double hidden layer model with (5,4) nodes is the best among all the training models.

Table 5. Neural network model training results

Parameter	Explanation	Parameter value	RMSE	MAE	Rsquare	Run time (min)
Hidden	Hidden layer node number	3	787.6142	243.4921	0.9328	19.77
Threshold	Training threshold	0.05				
Stepmax	Maximum training pace	6.00E+07				
Learningrate	Learning rate	0.0005				
Hidden	Hidden layer node number	4	1236.4231	324.5945	0.9265	12.04
Threshold	Training threshold	0.05				
Stepmax	Maximum training pace	6.00E+07				
Learningrate	Learning rate	0.0005				
Hidden	Hidden layer node number	5	nonconvergence	nonconvergence	nonconvergence	nonconvergence
Threshold	Training threshold	0.05				
Stepmax	Maximum training pace	6.00E+07				
Learningrate	Learning rate	0.0005				
Hidden	Hidden layer node number	(5,4)	681.4972	248.5108	0.9482	12.29
Threshold	Training threshold	0.05				
Stepmax	Maximum training pace	6.00E+07				
Learningrate	Learning rate	0.0005				

3.2.3. Random forest-based assessment models. In order to select the optimal number of leaf node splitting features, the experimental method was used to train the Mtry from 2 to 9 respectively with all other conditions remaining unchanged, and the test set data was used to validate the model accuracy, and in line with the previous paper, the root mean square error (RMSE), the mean absolute

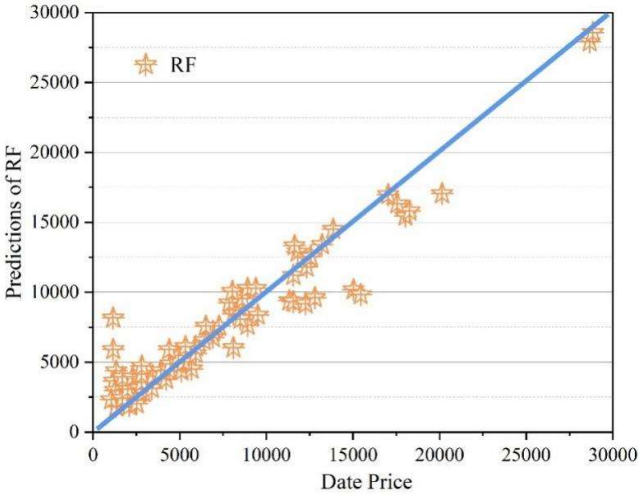
error (MAE), and goodness-of-fit (R^2) were computed and compared in order to select the optimal model. The random forest model training results are shown in Table 6. The training results show that as the number of node features Mtry increases, the variance explained and the goodness-of-fit (R^2) of the model generally show a trend of increasing and then decreasing, and the inflection point of the decrease occurs when the number of node features is seven. The root mean square error (RMSE), on the other hand, shows a change process of decreasing and then increasing, and the inflection point of the shift is also at the time when the number of nodal characteristics is 7. From the point of view of the mean absolute error (MAE), with the increase of the number of node features, the index is gradually decreasing, but it can be clearly found that before the number of node features is 7, the rate of decline is faster, and then tends to level off. After the above analysis, the overall performance of the model is the best when the number of node features Mtry is 7. Finally, the random forest model with 360 decision trees and 7 node features is chosen.

Table 6. Results of random forest model training

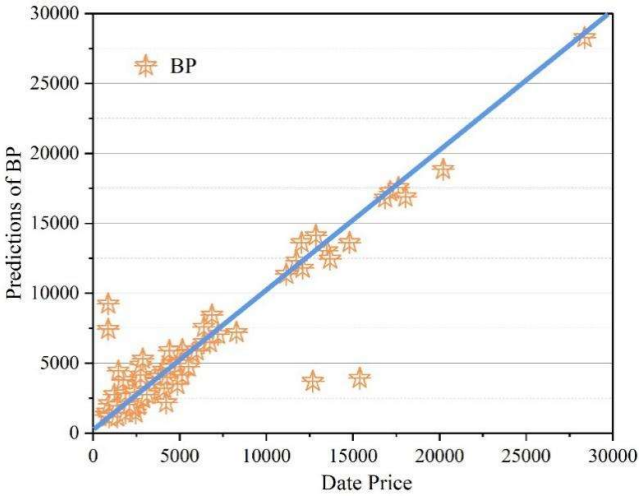
Node characteristic number	Variance interpretation	RMSE	MAE	Rsquare
2	0.8915	919.45	350.2441	0.897
3	0.9314	670.1803	200.1753	0.9481
4	0.9384	585.3155	157.1488	0.9618
5	0.9428	543.868	136.4869	0.9678
6	0.9413	530.1975	130.015	0.9697
7	0.9433	507.3682	121.4824	0.9727
8	0.9403	510.4915	121.0915	0.9723
9	0.9397	514.4275	118.5175	0.9718

3.3. Comparative analysis of assessment effects

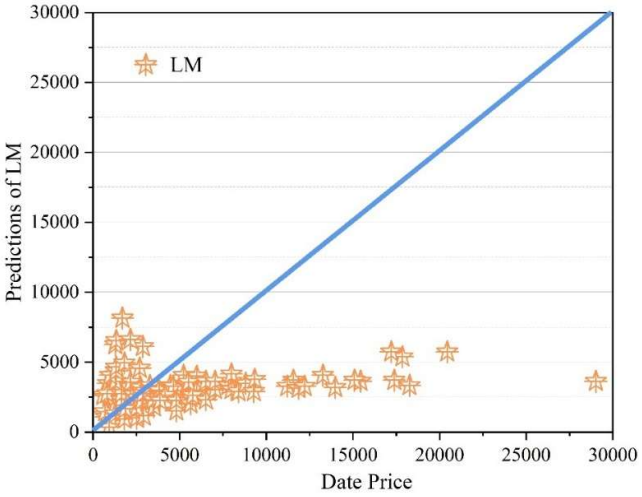
3.3.1. Comparison of the assessment effectiveness of the four models. Based on the above empirical analysis, organizing the assessment prediction effect of multiple linear regression model (LM), BP neural network model (BP), random forest model (RF) and Stacking model, this paper carves the relationship between the predicted value and the true value of the assessment of the random forest, Stacking model, neural network and multiple linear regression model. The relationship between the predicted value and the true value of the four models is shown in Fig. 5, and (a) to (d) show the relationship between the predicted value and the true value of the RF model, the BP model, the LM model and the Stacking model, respectively. It can be seen that the predicted values of the machine learning model are roughly distributed on the 45° line, indicating a better prediction effect. However, compared with the Stacking model, some of the predicted values of the neural network model are seriously deviated from the 45° line, indicating that the prediction effect of the stochastic Stacking model is still better than that of the neural network model. Finally, the predicted values of the multiple linear regression model deviate far from the true values, especially for the assessment results of high-value data assets, which deviate to a greater extent, and the overall prediction effect is poor. Overall, the predicted values of the Stacking model at different price points are more closely aligned with the true values, while the predicted results of the Random Forest Model model at various price points are not as comprehensive as those of the Stacking model. Therefore, the comparison of the evaluation effects of the four models shows that the Stacking model has the best prediction effect.


RF
Predictions of RF
Date Price
0
5000
10000
15000
20000
25000
30000

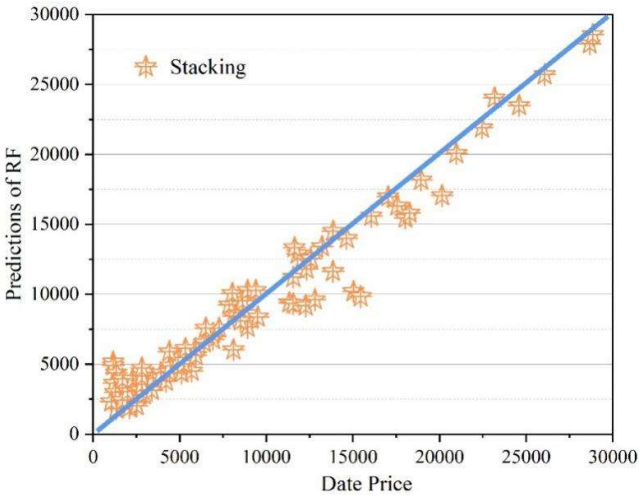
(a)RF


BP
Predictions of BP
Date Price
0
5000
10000
15000
20000
25000
30000

(b)BP


LM
Predictions of LM
Date Price
0
5000
10000
15000
20000
25000
30000

(c)LM



(d)Stacking

Fig. 5. The relationship between the 4 model predictions and the real value

A comparison of the evaluation effects of the four models is shown in Table 7. Taking all aspects together, the Stacking model has better performance in data asset value assessment. In terms of evaluation effect, the model constructed by using machine learning method is significantly better than the traditional multiple linear regression model, and the root mean square error (RMSE) and mean absolute error (MAE) of the machine learning model are much smaller than that of the multiple linear regression model, and at the same time, the goodness of fit is much larger than that of the multiple linear regression model. This also indicates that the effects of characteristics such as data capacity, data size, freshness and data quality on the value of data assets are indeed nonlinear, and thus machine learning models are more suitable for the assessment of data asset value. Looking at the machine learning models separately, the Stacking model outperforms the BP neural network model in all three metrics. It is worth mentioning that the root mean square error (RMSE) of the Stacking model is 498.1884, which is 202.0478 lower than that of the BP neural network model, and only about one-fifth of that of the multivariate linear regression model.The mean absolute errors (MAEs) of the Stacking model, the Random Forest model, the BP neural network model, and the multivariate linear regression model, respectively, are 119.6524, 173.5624, 264.5428 and 1177.718, which shows that Stacking model has the lowest MAE. Finally, the goodness of fit of the Stacking model is 97.12%, which is also higher than the other three models, showing a more accurate prediction.

Table 7. The evaluation effect of the 4 models is compared

Index	Multivariate linear regression model(LM)	BP neural network model(BP)	Random forest model(RF)	Stacking model
RMSE	2506.8924	700.2362	587.4372	498.1884
MAE	1177.718	264.5428	173.5624	119.6524
R^2	0.1928	0.9414	0.9635	0.9712

3.3.2. Comparison of metrics of different algorithmic models on the test set. The classification prediction index data of BP algorithm, LM algorithm, RF algorithm, and Stacking algorithm are shown in Table 8. The calculation shows that the overall accuracy of the enterprise data asset pricing prediction model based on Stacking algorithm is 96.06%, precision rate is 97.87%, recall rate is 97.75%, and F1 value is 0.9646%. In terms of the overall metrics, the metrics of the enterprise data asset pricing prediction model based on Stacking algorithm are all higher than the metrics of the other base models. Among the three base models, the RF algorithm-based enterprise data asset pricing

prediction model has the best prediction results with the highest of all four metrics, the BP algorithm is the second highest, and the LM algorithm is the worst.The prediction accuracy, precision, recall, and $F1$ values of the Stacking algorithm are higher than the three base models by 12.2%- 9.34%, 10.9%-9.54%, 12.22%-9.43%, and 11.92%-9.29%, respectively. Obviously, the prediction index of the Stacking algorithm obtained by integrating the three base models "BP algorithm, LM algorithm, and RF algorithm" increases significantly.

Table 8. Different algorithm models are compared in the test set

Algorithm name	Accuracy rate	Precision	Recall rate	F1 value
RF	0.8672	0.8833	0.8832	0.8717
BP	0.8545	0.8706	0.8724	0.8579
LM	0.8386	0.8697	0.8553	0.8454
Stacking	0.9606	0.9787	0.9775	0.9646

4. Conclusion

In order to evaluate enterprise data asset pricing, this paper constructs an integrated Stacking algorithm on the basis of multiple linear regression model, BP neural network model and random forest regression algorithm model to predict enterprise data asset pricing.

1) This paper obtains the absolute pricing element indicators of seven categories of experimental data "data size, scene coverage, time span, information entropy score, data sensitivity, data freshness, data citation, and data score", and explores and finds that there is no linear relationship between data scores and data products. The regression results indicate that the value of data assets with data capacity, data size, data quality and freshness are the main influencing factors affecting the pricing of data assets.

2) The results of RMSE, R^2 and MAE show that the prediction accuracy of the double hidden layer model is better than the single hidden layer model. And from the prediction accuracy of the model, it is found that the double hidden layer model with the number of nodes (5,4) is the best among all the trained models. Its RMSE, MAE, Rsquare and Run time (min) values are 681.4972, 248.5108, 0.9482 and 12.29min respectively. And the overall performance of the model is best when the number of node features is 7.

3) The performance of the enterprise data asset pricing prediction model based on the Stacking algorithm all outperforms the other three base models, with an increase in accuracy, precision, recall, and F1 value of 12.2%-9.34%, 10.9%-9.54%, 12.22% -9.43%, and 11.92%-9.29%.

Funding

1) The 2021 Research and Practice Project of Higher Education Reform in Henan Province "Internet +Accounting Factory" Research and Practice of Digital Intelligence Education Model in Vocational School under the Background of Integration of Production and Education "(Project Number: 2021SJGLX910).

2) The 2024 Research and Practice on Teaching Reform Project of Zhengzhou Vocational College Of Finance And Taxation. "Model Innovation, Resource Support, Mechanism Escort: Exploration and Practice of Accounting Intelligence Talent Cultivation in Higher Vocational Colleges" (Project Number: 2024JG01).

3) The 2024 Henan Provincial Education Science Planning Project of the Department of Education of Henan Province "Innovative Practice of Digital Intelligence Education Mode for Accounting Majors in Higher Vocational Colleges under the Background of New Quality Productivity" (Project Number: 2024YB0554).

References

[1] Sadowski, J. (2019). When data is capital: Datafication, accumulation, and extraction. Big data & society, 6(1), 2053951718820549.

[2] Cherrington, M., Lu, Z., Xu, Q., Thabtah, F., Airehrour, D., & Madanian, S. (2020). Digital Asset Management: New Opportunities from High Dimensional Data—A New Zealand Perspective. In Advances in Asset Management and Condition Monitoring: COMADEM 2019 (pp. 183-193). Cham: Springer International Publishing.

[3] Cheng, Q., Gong, Y., Qin, Y., Ao, X., & Li, Z. (2024). Secure Digital Asset Transactions: Integrating Distributed Ledger Technology with Safe AI Mechanisms. Academic Journal of Science and Technology, 9(3), 156-161.

[4] Birch, K., Cochrane, D. T., & Ward, C. (2021). Data as asset? The measurement, governance, and valuation of digital personal data by Big Tech. Big Data & Society, 8(1), 20539517211017308.

[5] Pei, J., Zhu, F., Cong, Z., Luo, X., Liu, H., & Mu, X. (2021, August). Data pricing and data asset governance in the AI era. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 4058-4059).

[6] Wang, T., Jiang, L., & Wang, X. (2022, December). Research on the Pricing Method of Power Data Assets. In International Conference on Big Data (pp. 52-60). Cham: Springer International Publishing.

[7] Alkhard, A. (2024). Enhancing asset management through integrated facilities data, digital asset management, and metadata strategies. Construction Economics and Building, 24(3), 76-94.

[8] Hao, J., Yuan, J., Li, J., Liu, M., & Liu, Y. (2023). Ensemble pricing model for data assets with ranking-pruning-averaging strategy. Procedia Computer Science, 221, 813-820.

[9] Nolin, J. M. (2020). Data as oil, infrastructure or asset? Three metaphors of data as economic value. Journal of Information, Communication and Ethics in Society, 18(1), 28-43.

[10] Birch, K. (2023). Data Assets. In Data Enclaves (pp. 41-59). Cham: Springer Nature Switzerland.

[11] Abdirad, H., & Dossick, C. S. (2020). Rebaselining asset data for existing facilities and infrastructure. Journal of Computing in Civil Engineering, 34(1), 05019004.

[12] Ding, C., Wen, K., & Tian, Y. (2024, August). Research on Pricing Methods of Railway Data Assets in Internal and External Market. In International Conference on Traffic and Transportation Studies (pp. 481-487). Singapore: Springer Nature Singapore.

[13] Veldkamp, L. (2023). Valuing data as an asset. Review of Finance, 27(5), 1545-1562.

[14] Li, Z., Ni, Y., Yang, L., & Gao, Y. (2019, March). Review and prospect of data asset research: Based on social network analysis method. In 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA) (pp. 345-350). IEEE.

[15] Hannila, H., Silvola, R., Harkonen, J., & Haapasalo, H. (2022). Data-driven begins with DATA; potential of data assets. Journal of Computer Information Systems, 62(1), 29-38.

[16] Telenti, A., & Jiang, X. (2020). Treating medical data as a durable asset. Nature Genetics, 52(10), 1005-1010.

[17] Eichler, R., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2022, July). From data asset to data product–the role of the data provider in the enterprise data marketplace. In Symposium and Summer School on Service-Oriented Computing (pp. 119-138). Cham: Springer International Publishing.

[18] Zhao, Y. (2024). The Influence Mechanism of Data Asset Management Regulation and New Quality Productivity on Innovation. Journal of Statistics and Economics (ISSN: 3005-5733), 1(1), 35.

[19] Nixon, J., & Pena, E. (2019, April). The evolution of asset management: Harnessing digitalization and data analytics. In Offshore Technology Conference (p. D021S016R005). OTC.

[20] Wang, J., Li, Y. S., Song, W., & Li, A. H. (2018). Research on the theory and method of grid data asset management. Procedia computer science, 139, 440-447.

[21] Collins, V., & Lanz, J. (2019). Managing data as an asset. The CPA Journal, 89(6), 22-27.

[22] Gavrikova, E., Volkova, I., & Burda, Y. (2022). Implementing asset data management in power companies. International Journal of Quality & Reliability Management, 39(2), 588-611.

[23] Li, Y. (2024). Concepts, Accounting Treatment and Pricing of Data Assets. Journal of Economic Insights, 1(1).

[24] Yanlin, W., & Haijun, Z. (2020, May). Data asset value assessment literature review and prospect. In Journal of Physics: Conference Series (Vol. 1550, No. 3, p. 032133). IOP Publishing.

[25] Badewitz, W., Hengesbach, C., & Weinhardt, C. (2022, June). Challenges of pricing data assets: a literature review. In 2022 IEEE 24th Conference on Business Informatics (CBI) (Vol. 1, pp. 80-89). IEEE.

[26] Fernandez, R. C., Subramaniam, P., & Franklin, M. J. (2020). Data market platforms: trading data assets to solve data problems. Proceedings of the VLDB Endowment, 13(12), 1933-1947.

[27] You, J., Lou, S., Mao, R., & Xu, T. (2022). An improved FMEA quality risk assessment framework for enterprise data assets. Journal of Digital Economy, 1(3), 141-152.

[28] Li, Y., Luo, C., Dong, L., & Gui, M. (2022). Data asset disclosure and nonprofessional investor judgment: Evidence from questionnaire experiments. Mobile Information Systems, 2022(1), 8116063.

[29] Zhao Qi. (2022). Modeling and Analysis Method of National Fitness Big Data for Basketball Projects Based on a Multivariate Statistical Model. Security and Communication Networks

[30] Jeonghoon Choi & Dongjun Suh. (2025). A depthwise convolutional neural network model based on active contour for multi-defect wafer map pattern classification. Engineering Applications of Artificial Intelligence(PB),109707-109707.

[31] Ahmed Alshembari,Markus Kronenberger,Sophie Burgmann,Katja Schladitz & Wolfgang Breit. (2024). Separation of sand and gravel particles in volume images using a random forest. Materials Today Communications110957-110957.

[32] Zhaowei Jie,Xiaohan Zhu,Hanyu Zhang,Hanyang Zheng,Can Hu,Zhanfang Liu... & Hongcheng Mei. (2024). A novel method for tracing gasoline using GC-IRMS and Relief-Stacking fusion model. Microchemical Journal112081-112081.

[33] Chong Hu, Rui Deng, Yuanpeng Zhang, Jie Chen, Ming Li, Lixia Dang... & Kun Xiong. (2023). Lithology identification technology based on the stacking fusion model. International Journal of Oil, Gas and Coal Technology(4),337-358.