Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-06016-y OPEN

How to price a dataset: a deep learning framework for data monetization with alternative data

Jun Hao[1,2]✉, Zeyu Deng[1,2], Jin Li[3] & Jianping Li[1,2]✉

In the context of the digital economy, data is regarded as a critical production factor, and its pricing is essential for promoting the circulation of data assets, ensuring transaction fairness, and driving data sharing and economic development. In this study, an intelligent data pricing model is proposed, which integrates traditional numerical features with non-traditional textual information to improve pricing accuracy. Traditional pricing models often fail to fully capture the complete value of data assets, particularly by overlooking alternative data such as functional descriptions or textual metadata associated with the data. To address this issue, this paper develops a deep learning pricing framework that leverages the light gradient boosting machine (LGBM) and bidirectional encoder representations from transformers (BERT) for textual analysis, significantly improving pricing precision. Experimental results demonstrate that the proposed model achieves lower prediction errors across various training-test ratios, reducing pricing errors by 63.5% compared to traditional models. These findings validate the important role of non-traditional data in enhancing the accuracy of data asset pricing. Consequently, this study enhances the effectiveness of data valuation and underscores the value of alternative data sources in data pricing. It further highlights the potential benefits of incorporating additional data sources and deep learning techniques to enhance the performance of pricing models in future research.

[1] University of Chinese Academy of Sciences, Beijing, China. [2] MOE Social Science Laboratory of Digital Economic Forecasts and Policy Simulation at UCAS, Beijing, China. [3] Xi'an Jiaotong University, Xi'an, China. ✉email: haojun@ucas.ac.cn; ljp@ucas.ac.cn

Introduction

In the rapidly evolving landscape of the digital economy, data has emerged as a new production factor, often referred to as the oil of the 21st century (Veldkamp, 2023). With the swift advancements in the internet and big data technologies, the collection, storage, processing, and analysis of data have become unprecedentedly efficient and accessible. This has enabled the extensive application and realization of data's value across various fields. From corporate market analysis to governmental public policy formulation, from personal health management to social science research, data plays an increasingly pivotal role. Accurate pricing not only facilitates the circulation of data assets but also ensures fair and transparent transactions between data providers and users. Moreover, a reasonable data pricing mechanism is crucial for the successful implementation of data-sharing and open data initiatives, helping to unlock the social value of data and fostering innovation and economic development. Consequently, the challenge of how to price data assets reasonably has become both urgent and complex.

Currently, scholars have conducted extensive research on data asset pricing models from various perspectives, proposing a range of different pricing models such as game theory-based pricing, auction-based pricing, information entropy-based pricing, and data feature-based pricing models (Jun et al. 2023a). X. Zhang et al. (2023) developed a game theory model to achieve data asset pricing and further explored the impact of data monetization on inter-firm competition. Yu and Zhang (2017) established a bi-level programming valuation model that incorporates data quality. To address the limitations of single pricing models in fully capturing the overall patterns of data asset prices. Jun et al. (2023b) proposed a novel integrated data asset pricing framework. This model effectively integrates pricing results by generating ten machine learning pricing models and introducing a ranked pruning average strategy. Abbasi et al. (2023) introduced a blockchain-based industrial data trading system that ensures the security and transparency of data transactions, offering advantages over traditional systems. Mehta et al. (2021) introduced the concept of ideal records and developed a utility framework to determine appropriate pricing strategies for both data buyers and sellers. A tractable model was established, successfully leveraging its unique structure to obtain optimal and near-optimal data sales mechanisms.

However, despite substantial empirical evidence and stylized facts verifying the significant impact of data quality, accuracy, and timeliness on data pricing, existing data asset pricing models have yet to consider alternative data such as functional descriptions of data. Incorporating alternative data, such as textual information, in addition to conventional features, may be an effective solution for describing data characteristics and enhancing the accuracy of data pricing (Gao et al. 2020). Alternative data can provide a more comprehensive analytical perspective, aiding businesses and investors in making more informed decisions (Liu et al. 2023). The applicability of alternative data has already been demonstrated in various disciplines, including financial markets (Lehrer et al. 2021; Sheng et al. 2024). Despite its promising potential, the application of alternative data in data asset pricing remains at a relatively early stage.

Our research further explores the untapped potential of alternative data by introducing functional descriptions and other textual data to enhance the effectiveness of pricing models. Advanced machine learning algorithms, specifically LGBM, were employed to develop an efficient and accurate data asset pricing model. This improved pricing method can assist data enterprises and data brokers in conducting rapid and efficient data asset pricing, thereby ensuring the rationality and transparency of data asset transactions. Additionally, the BERT text analysis algorithm was introduced to effectively extract, interpret, and integrate textual information. A set of machine learning algorithms was utilized to train and learn the data pricing model using multi-modal data. Ablation experiments were designed using the controlled variable method to validate the pricing effectiveness of the model.

The potential contributions of this study are threefold. Firstly, a novel data asset pricing framework is proposed, in which textual and numerical features are jointly exploited to enhance the model's ability to capture diverse pricing determinants. Empirical results demonstrate its superior predictive performance over baseline models, with the LGBM-based implementation achieving a mean squared error (MSE) as low as 0.9941. This confirms the framework's superiority in capturing pricing-relevant patterns that traditional numerical features fail to represent. Secondly, alternative textual attributes, particularly functional descriptions, are integrated into the pricing model. When encoded via BERT and fused with numerical inputs, these features significantly improve prediction accuracy, as verified through ablation experiments and scenario-based analyses. This underscores the unique value signals embedded in unstructured textual data. Thirdly, a scenario-based approach is introduced to estimate feature importance by combining SHAP attribution and rank-based exclusion. Results show that high-value features (e.g., data descriptions and size) are crucial for accuracy, while certain low-value features may introduce noise. This method provides actionable guidance for feature selection and data valuation in practice by quantifying the marginal utility of individual attributes.

The structure of this paper is organized as follows: Section "Literature review" provides a literature review covering data pricing models, alternative data, and data value measurement. Section "Methods" presents a detailed introduction to the proposed pricing framework. Section "Experimental evaluation" outlines the experimental setup and evaluation process. Section "Result" analyzes the experimental results. Section "Scenario-based data value estimation" delves into the value analysis of driving factors. Finally, the section "Conclusion" offers the conclusion and future outlook of the study.

Literature review

Our research primarily pertains to the monetization of data commodities or assets, with a focus on three intertwined themes: data pricing methodologies, the use of alternative data in financial valuation, and data value analysis techniques. Accordingly, each of these aspects will now be reviewed in detail.

Data pricing. With the advent of the information age, research in data asset pricing has gradually emerged, yet it remains relatively underdeveloped (M. Zhang et al. 2023). Early studies primarily focused on the refinement of traditional asset pricing theories, attempting to apply existing financial asset pricing models to data assets (Anand and Jha, 2022; Gu et al. 2021). As data economics and business intelligence have advanced, increasing attention has been directed toward the distinctive value assessment and pricing methods specific to data (Xu et al. 2022).

Currently, data pricing methods can be categorized into two main types: economic theory-based and computational intelligence-based approaches (Jun et al. 2023a). Relevant studies have systematically reviewed existing data pricing methodologies, with detailed information provided in the literature (Jun et al. 2023a). Economic theory-based methods include cost-based pricing, consumer perceived value pricing, supply and demand-based pricing, differential pricing, dynamic pricing, game theory-

based pricing, and auction pricing (Henry et al. 2023; Liang and Yuan, 2021; Mumbower and Garrow, 2014; Yu and Zhang, 2017). Gneezy et al. (2014) posited that data quality is correlated with price. Such studies often classify data quality based on different dimensions. For example, Yang et al. (2019) and Yu and Zhang (2017) assessed data quality, constructed consumer utility functions, and priced data based on consumer utility and inherent value. Game theory-based pricing methods focus on the decision-making processes and equilibrium issues arising from the direct interactions between data providers and data consumers (Xiao et al. 2021). Commonly used approaches include non-cooperative games (Nguyen Cong et al. 2016) and Stackelberg games (Liu et al. 2019). Auctions, as a significant application of incomplete information games, are frequently employed in data pricing, such as double auctions and sealed-bid auctions (Cao et al. 2017). Additionally, pricing methods that consider the dynamics of supply and demand of the data market are also utilized (Mehta et al. 2021).

Pricing methods based on computational intelligence encompass query pricing, privacy computation, federated learning, ensemble learning, and more (Feng et al. 2023; Hao et al. 2024; Peyvandi et al. 2022; Zhang et al. 2022). Koutris et al. (2015) implemented automated pricing for arbitrary queries by presetting view prices. Subsequently, Miao et al. (2022) explored the pricing of incomplete data queries. A comprehensive pricing mechanism and two novel price functions—usage and completeness-aware price function, and quality, usage, and completeness-aware price function—were proposed, considering data contribution/usage, data integrity, and query quality. Tian et al. (2022) investigated data boundary pricing based on the Shapley value, addressing the three-party data market pricing problem involving data owners, model purchasers, and platforms. Niu et al. (2022) proposed a contextual dynamic pricing mechanism with reserve price constraints. Zhang et al. (2022) designed incentive schemes within a federated learning framework to achieve data pricing. Xu et al. (2017) examined the pricing issue for data collectors purchasing data sequentially from multiple data owners, modeling the collector's sequential decision-making problem as a multi-armed bandit problem.

In particular, machine learning-based models such as decision trees, neural networks, and ensemble frameworks have demonstrated strong potential for capturing complex feature-price mappings in high-dimensional data spaces. These models facilitate global optimization and support the integration of heterogeneous data sources, making them especially suitable for modern data marketplaces where structured and unstructured features coexist.

Data pricing, as a critical business strategy, involves determining the price of products or services based on data characteristics, market demand, and competitive conditions. However, with market dynamics and increasing uncertainty in customer behavior, traditional pricing methods may fall short in meeting the demands of fast-evolving business environments, thereby impacting the accuracy and flexibility of pricing strategies.

The application of alternative data in financial issues. Alternative data, distinct from traditional financial data, is increasingly adopted to enhance market understanding, especially in contexts requiring real-time insights and behavioral indicators. By capturing micro-level dynamics and shifts in investor sentiment, it supplements conventional data sources and enables more accurate risk assessments and value discovery (Ma et al. 2025).

Recent literature has explored its applications across financial domains. For example, Hansen and Borch (2022) highlighted social media sentiment's role in stock price prediction, demonstrating how unstructured textual signals can improve valuation models. Zhang et al. (2024) showed that convolutional neural networks, when incorporating pandemic-related indicators, maintained strong predictive performance during crises. Dessaint et al. (2024) found that access to short-term-oriented alternative data shifted analysts' forecasting horizons, enhancing valuation relevance.

Some studies also apply alternative data to credit scoring and fintech lending (Djeundje et al. (2021); Hlongwane et al. (2024); Tigges et al. (2024)), but these are domain-specific and do not address generalizable pricing mechanisms for data assets.

Despite growing interest, most existing research emphasizes either predictive tasks or sector-specific outcomes, without integrating alternative data into unified pricing frameworks. Traditional methods, reliant on static models or historical records, often overlook the full value of unstructured data. Moreover, limited work explores how structured and unstructured data can be fused into multimodal representations suitable for asset pricing.

Therefore, a clear research gap exists in designing pricing models that jointly leverage numerical and textual inputs to assess data value. Addressing this gap is essential for improving the effectiveness of data valuation in dynamic, information-rich environments such as financial markets. This study responds to this gap by proposing a multimodal framework combining machine learning and text analysis to advance data asset pricing.

Data value analysis. In the era of big data, the importance of measuring data value has become increasingly prominent. Due to the vast volume and diversity of data, the value density of data tends to be low, which not only escalates processing and handling costs but also poses significant challenges to data-driven modeling tasks. Low-quality data often leads models to learn complex or irrelevant patterns, resulting in decreased accuracy and learning efficiency. High-quality data input is crucial for the success of a model, and measuring data value serves as an essential tool for assessing data quality. By measuring data value, beneficial information for model construction can be identified, while irrelevant or low-value noise can be excluded, thereby enhancing the model's predictive and generalization capabilities. Assessing data of varying quality in practical scenarios allows for informed decisions regarding the retention or elimination of data based on its quality or value. Thus, accurately measuring data value aids enterprises and individuals in swiftly identifying valuable information within massive datasets, thereby improving data utilization efficiency.

However, current research on data value measurement remains insufficient. Most existing methods for assessing dataset quality lack a unified quantitative measurement standard and fail to incorporate supervisory mechanisms from real-world model outputs. Scholars have attempted to characterize data quality using metrics such as Shannon entropy (Li et al. 2022; Xu et al. 2021) and signal-to-noise ratio (Yu et al. 2022), and have explained data value through the principle of uncertainty reduction (Kikuchi and Kronprasert, 2012). Nevertheless, these metrics have not effectively revealed the mechanisms of value creation at the data level. In this context, the leave-one-out (LOO) method can effectively evaluate the marginal utility of individual data points relative to the entire dataset. However, its capability to assess the cumulative value of data is limited, thus constraining its broader applicability in big data analytics.

Given the complex interrelationships among data, the SHAP (Shapley Additive Explanation) method has been introduced as an innovative tool for data value assessment. The SHAP method, which draws on the core principles of game theory, is a post-hoc

interpretability technique. It precisely measures the impact of each feature and its interactions on the model output by calculating the marginal contribution values, known as Shapley values, for all input features and their interactions within the model. This method not only ranks features based on their contribution levels, helping to identify those with the greatest impact on model predictions, but also adheres to the principle of fair allocation inherent in Shapley values from cooperative game theory. This ensures that each feature's contribution to the model output is appropriately recognized. The inherent fairness of the SHAP method makes its interpretative results more objective and impartial, aiding in the avoidance of potential biases in the interpretation outcomes.

Although existing studies have made progress in data pricing, the use of alternative data, and data value analysis, several key gaps remain. First, many models still treat structured and unstructured data separately, lacking a unified approach that integrates numerical and textual information for pricing. Second, there is limited research on pricing specific types of data assets, such as transaction records that include descriptive attributes. Third, while machine learning is used in pricing tasks, few models are designed specifically for data assets or make full use of interpretable text features. These limitations point to the need for a dedicated multimodal pricing framework, which this study proposes to address.

Methods

Problem formulation. The task of data asset pricing fundamentally involves identifying patterns from complex observational samples to estimate the relationship between input features and asset price. In our setting, let $x_i \in \mathbb{R}^d$ denote the i-th data asset instance, composed of structured and unstructured features, and let $y_i \in \mathbb{R}$ represent its corresponding price. Denote the input space as $X = \{x_{i1}, x_{i2}, \cdots, x_{id} | i\} \subseteq R^d$, and the label space as $Y = \{y_i | i\} \subseteq R$. $P(X)$ denotes an unknown probability distribution over X. The goal is to learn a pricing function $f : X \to Y$, given a training dataset $D = \{(x_i, y_i)\}_{i=1}^N \subseteq (X, Y)^N$.

Assuming that the mapping from X to Y lies within a hypothesis space H, the core task of data asset pricing is to learn a function $\hat{f} \in H$ from the training data D, such that the expected error between predicted and actual values is minimized:

$$\hat{f}(x_i) = \arg\min_{\hat{f}(x_i)} E_{x_i, y_i}(L(y_i, \hat{f}(x_i))) \tag{1}$$

Here, $L(y_i, \hat{f}(x_i))$ denotes the loss function. A smaller loss indicates a more accurate pricing model. In this study, we adopt the mean squared error (MSE) as the loss function.

Given the unknown and potentially nonlinear nature of the mapping between data features and prices, supervised learning models offer a principled approach to data pricing. By learning from labeled samples, they approximate the latent value function f that maps heterogeneous inputs—both structured and unstructured—to scalar price estimates. This allows the model to capture complex valuation patterns and support scalable, accurate pricing across diverse data assets.

Pricing framework. To better address the complexity of data asset valuation, we propose a multimodal pricing framework. It integrates structured indicators with unstructured textual descriptions to reflect the multifaceted nature of data value. As shown in Fig. 1, the framework consists of three main stages: data asset feature construction, integrated pricing, and scenario-based value assessment.

Data asset feature representation. In the first stage, both numerical and textual information are extracted to form a rich feature space for each data asset. The numerical side includes standard metadata such as data volume, redundancy, and rarity, which are quantified through statistical analysis. Meanwhile, the textual side focuses on functional descriptions (e.g., the content descriptions, use scenarios, and target consumers), which are processed using a pretrained language model (BERT) to generate dense vector representations. These heterogeneous features are integrated into a unified representation space, enabling the model to jointly leverage structured indicators and semantic attributes for pricing.

Integrated pricing module. Once features are constructed, the framework feeds them into a unified pricing module based on ensemble learning. Specifically, a LightGBM-based regressor is trained using both structured and unstructured features, and it performs global optimization by aggregating results from multiple locally learned regression trees. The model dynamically learns the mapping between multimodal features and observed prices, capturing complex interactions that traditional rule-based or linear models may miss. Furthermore, internal tracking of feature importance enhances interpretability and supports more accurate data pricing.

Scenario-based value assessment. To assess how different features influence pricing, the framework includes a scenario-aware evaluation component. Text features are first aggregated using multilayer perceptrons to preserve semantic information. Then, a ranking-based feature ablation is applied—features are sorted by importance and iteratively removed to assess their effect on model performance. This sensitivity analysis helps uncover which modalities and specific attributes contribute most to pricing accuracy, and provides actionable insights into the composition of data value.

Compared with traditional pricing frameworks, our design directly addresses two major challenges in data asset pricing: (i) how to systematically incorporate alternative (textual) data into the pricing process and (ii) how to model complex feature interactions without relying solely on handcrafted rules. This ensures that the framework is not only predictive, but also interpretable and adaptable to varied pricing scenarios.

Bidirectional encoder representations from transformers. BERT is a pretrained language model designed to capture contextual semantics through a bidirectional Transformer encoder architecture. It computes dependencies between tokens using a multi-head self-attention mechanism, enabling it to learn rich semantic representations from unlabeled text corpora (Alonso-Robisco and Carbo, 2023). Its overall architecture, as depicted in Fig. 2, consists of an input layer, an encoding layer, and an output layer.

Each transformer block employs a multi-head attention mechanism, where embedded inputs are linearly transformed into Query, Key, and Value vectors. The structure of a typical transformer encoder is illustrated in Fig. 3, and its core computations are defined by Eqs. (2) to (4).

$$\text{Attention}(Q, K, V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

$$head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

$$MultiHead(Q, K, V)Concat(head_1, \cdots, head_h)W^o \tag{4}$$

where the matrices W_i^Q, W_i^K, and W_i^V are the respective weight matrices for Q, K, and V. The variable (i) denotes the number of

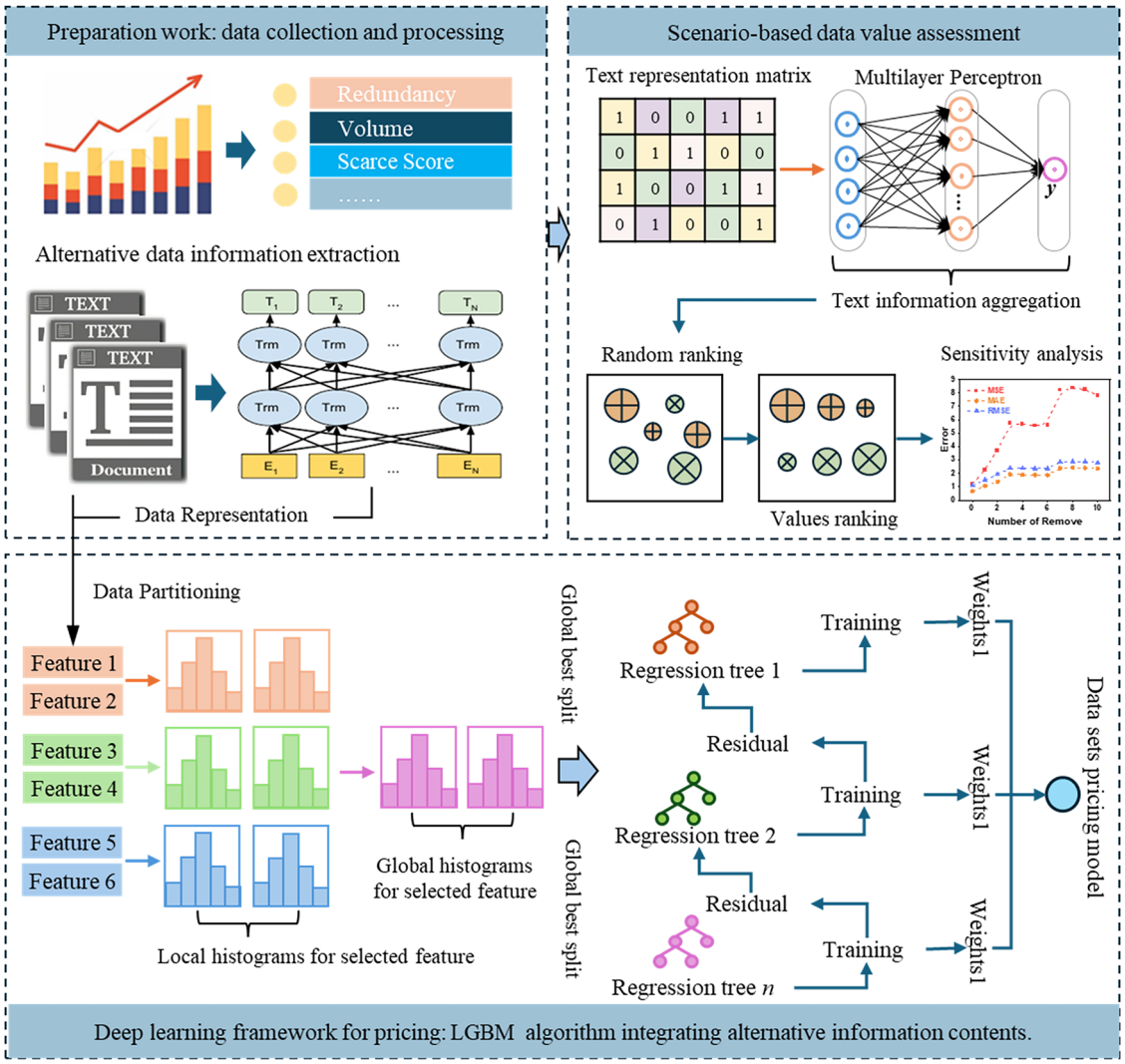

Fig. 1 Proposed framework.

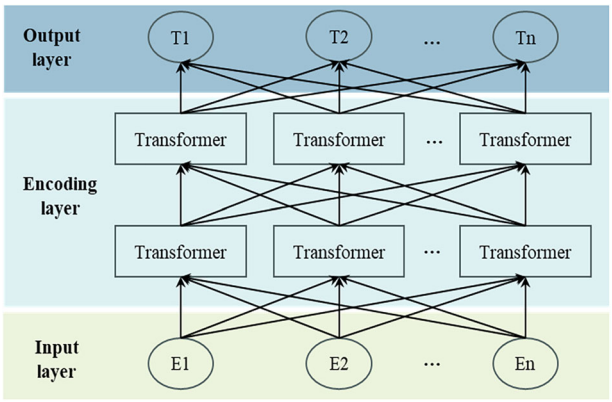

Fig. 2 Framework of BERT.

attention heads, W^o represents the mapping matrices of the multi-head attention mechanism.

In this study, BERT is applied to encode unstructured textual attributes of data assets—such as descriptions, functions, and usage contexts—into dense semantic vectors. These representations are later fused with structured indicators to form a multimodal input for the pricing model. The adoption of BERT allows the proposed framework to extract fine-grained contextual signals from domain-specific text, which enhances the model's capacity to assess data value beyond what numeric indicators alone can reveal.

LGBM. LightGBM is an efficient gradient boosting framework based on decision trees (Sheng et al. 2024). It improves training speed and reduces memory usage by using histogram-based algorithms and gradient-based one-sided sampling (J. Hao et al. 2023; Yuan et al. 2023). These design choices allow it to scale effectively to large datasets with high-dimensional features.

The LightGBM algorithm is expressed as a summation function composed of k base models, as shown in Eq. (5).

$$\hat{y}_i = \sum_{t=1}^{k} f_t(x_i) \tag{5}$$

where x_i denotes the input features of the i sample, f_t represents the t base model, and $\hat{y}_i$ signifies the predicted value of the i sample. The loss function can be expressed in terms of the predicted and true values, as shown in Eq. (6).

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) \tag{6}$$

where n represents the sample size, l denotes the loss function of the i sample, and y_i indicates the true value of the i sample. Based on these definitions, the objective function is formulated as

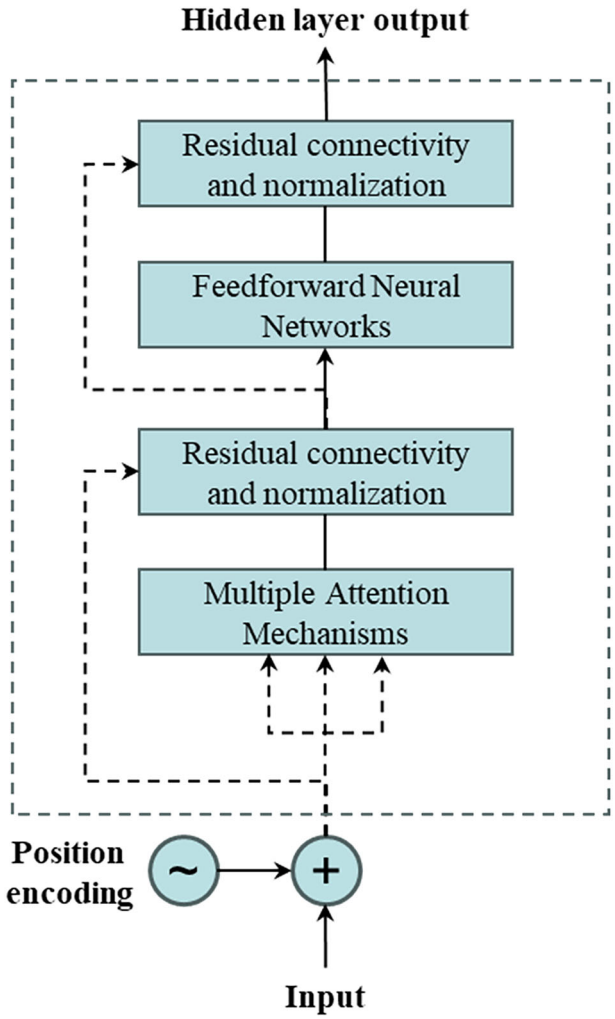

Fig. 3 Transformer encoder.

shown in Eq. (7).

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{k} \Omega(f_t) \tag{7}$$

where Ω represents the regularization term, and θ denotes the model parameters.

In this framework, LightGBM is employed to learn the mapping between multimodal features (textual and numerical) and the asset price. It is particularly well-suited for this task due to its ability to: (i) capture complex nonlinear relationships without requiring extensive manual feature engineering;

(ii) handle both sparse and dense features efficiently; and (iii) provide feature importance scores that enhance downstream interpretability (as utilized in section "LGBM"). By integrating LightGBM with the fused representations described in the section "Problem formulation", the proposed model enables precise and scalable valuation of data assets in heterogeneous information environments.

SHAP. SHAP (SHapley additive exPlanations) is introduced to measure the importance of different data features. SHAP provides a unified framework that uses Shapley values to explain the predictive behavior of machine learning models. Shapley values, derived from cooperative game theory, represent the average marginal contribution of each feature (Jiang et al. 2024). Specifically, SHAP assigns importance values to each predictive feature based on an additive feature attribution method that adheres to a set of desirable theoretical properties: local accuracy, missingness, and consistency. The explanatory model associated with this method is shown in Eq. (8):

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{8}$$

where $z' \in \{0,1\}^M$ is defined as a joint vector indicating whether the i feature is present (=1) or absent (=0). M represents the total number of features, and $\phi_i \in \mathbb{R}$. In this approach, the explanatory model employs a simplified input $x\prime$, which is mapped to the original input x via the mapping function $h_x(x\prime) = x$. This ensures that $g(z\prime) \approx f(h_x(z\prime))$, where $z\prime \approx x\prime$. The explanatory model is then established and trained to interpret the original model f. Given that SHAP measures feature importance by comparing the differences in model predictions with and without the feature, the SHAP value ϕ_i (i.e., the feature importance value) is calculated as the weighted average of all possible differences. This is expressed in Eq. (9):

$$\phi_i = \sum_{S \subseteq N\{x_i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_x(S \cup \{x_i\}) - f_x(S)] \tag{9}$$

where $N\{x_i\}$ denote the set of features excluding x_i, and S represent the subset of features excluding the i feature. The predictions of the model with and without the i feature are denoted by $f_x(S \cup \{x_i\})$ and $f_x(S)$, respectively. It should be noted that SHAP becomes an additive feature attribution method only if $\phi_0 = f_x(\phi)$.

In the proposed framework, SHAP is applied to the trained LightGBM model to attribute the predicted price of each data sample to its multimodal features, with the prediction value expressed as the sum of individual SHAP values. This enables transparent interpretation of feature-level contributions and supports informed decision-making in data asset pricing.

Experimental evaluation

Data. A dataset comprising 1426 entries was obtained from a data trading platform to evaluate the effectiveness of the pricing model. Each sample includes 12 numerical feature variables, such as product rating, usage frequency, data size, consistency, and structure. Descriptive statistics for these features are presented in Table 1. Additionally, the dataset contains textual information detailing the data, including descriptions, target audience, and functionality. To provide a clearer illustration, an example of an agricultural product data sample is shown in Fig. 4. As shown in Table 1, a skewness greater than 1 indicates a highly skewed distribution, specifically positive or right-skewed. This characteristic is further confirmed by Fig. 5. Consequently, a Box-Cox transformation (Box and Cox, 1964) was applied to normalize the data assets, as illustrated in Fig. 5.

To incorporate unstructured textual information into the pricing model, we employed the pretrained BERT-base-Chinese model from HuggingFace Transformers. Each data full textual description, including its name, function, and usage context, was tokenized using the associated BERT tokenizer with a maximum sequence length of 512. Longer sequences were truncated and shorter ones padded automatically. For each text instance, the final-layer [CLS] token embedding (768-dimensional) was extracted as a fixed-length semantic representation. The pretrained BERT model was used in a frozen configuration to extract contextual embeddings, without further fine-tuning. These embeddings were subsequently fused with structured numerical features to construct multimodal inputs for the pricing model.

For downstream analysis purposes (section "Scenario-based data value estimation"), the textual input was also decomposed

Table 1 Descriptive statistics.

	Max	Min	Media	Std	Ave	Skewness
Price	2989	10	291.5	764.95	624.99	1.2625
Goods score	5.00	3.00	4.00	0.4861	0.3610	−0.3881
Use num	230.00	0.00	26.00	23.3891	15.4926	3.1906
Size	1016.00	1.00	35.00	212.2704	149.6175	2.1857
Scarce score	5.00	3.00	4.00	0.7499	0.5672	−0.0183
Consistent score	5.00	3.00	4.00	0.7449	0.5702	−0.0554
Application score	5.00	3.00	4.00	0.7408	0.5561	0.0257
Structure score	5.00	3.00	4.00	0.7615	0.5814	0.0059
Dv score	5.00	3.00	4.00	0.7658	0.6046	−0.0750
Redundant score	5.00	3.00	4.00	0.7517	0.5776	0.0472
Integrality score	5.00	3.00	4.00	0.7615	0.6049	−0.1016
Tl score	5.00	3.00	4.00	0.7324	0.5412	0.0163
Category	8.00	1.00	5.00	1.8290	1.5565	0.1374

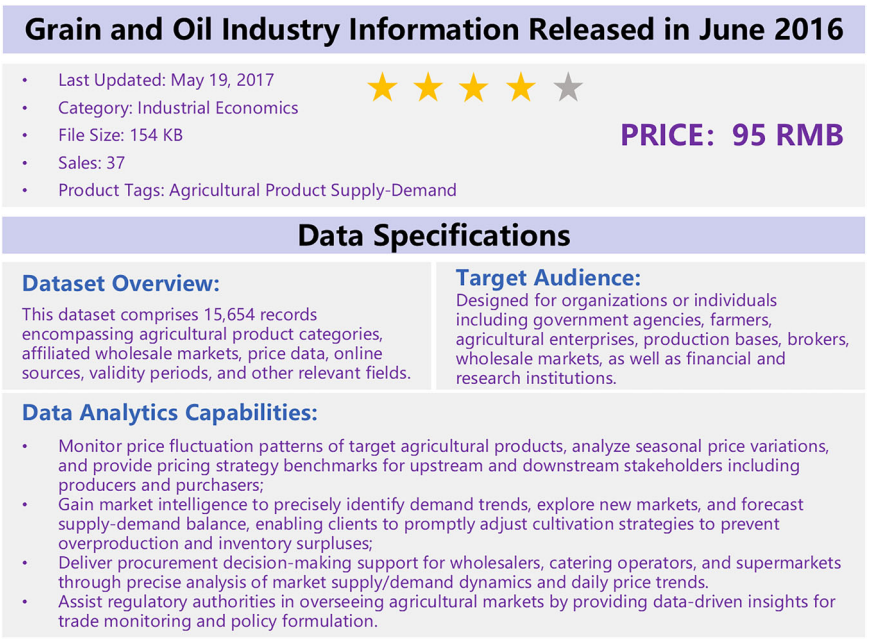

Fig. 4 Data sample description.

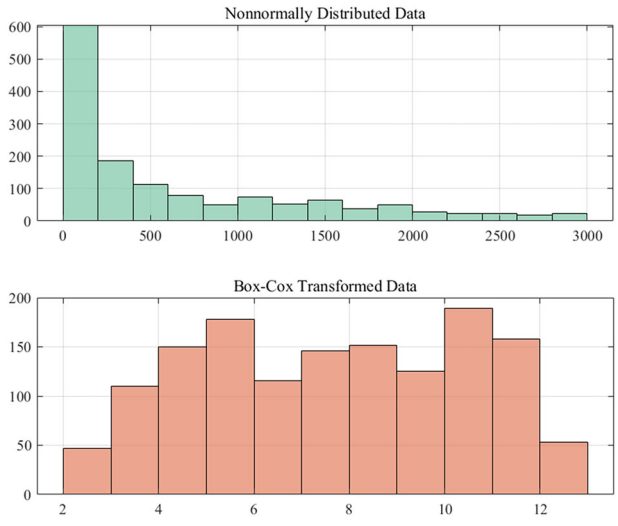

Fig. 5 Distribution of data price.

into four semantic subcomponents—title, description, target users, and function, and encoded separately using the same approach to evaluate their individual contributions to pricing performance.

Benchmarks and parameter setting. To ensure fair and comprehensive benchmarking, we selected representative models from various learning paradigms, including linear regression (MLR, Lasso), tree-based methods (DT, RF, GBDT, LGBM), kernel-based models (SVR), instance-based learners (KNN), and neural networks (MLP, LSTM). This diverse set covers both traditional and modern approaches, enabling evaluation across models with varying capacities for nonlinearity, interpretability, and data structure adaptability. The benchmark models and their corresponding parameters utilized in this study are detailed in Table 2.

Parameter settings were primarily drawn from default values recommended in the literature or official toolkits, with slight adjustments to ensure convergence. For instance, MLP uses two hidden layers (30, 50) with a 0.001 learning rate, while tree-based models like RF and GBDT use 100 estimators. SVR applies a

Table 2 Benchmarks and parameter settings.

NO.	Model	Parameter setting
1	MLP	Hidden layer sizes = (30, 50); Activation = "ReLU"; Solver = "Adam"; Max iterations = 500; Alpha = 0.001; Learning rate = "Constant"; Learning rate initial = 0.001
2	MLR	No parameters needed
3	Lasso	Alpha = 0.1
4	DT	Random state = 0
5	SVR	Kernel = "RBF"; C = 1.0; Epsilon = 0.1
6	RF	Number of estimators = 100
7	GBDT	Number of estimators = 100; Learning rate = 0.1; Max depth = 3
8	KNN	Number of neighbors = 5
9	LSTM	Number of hidden units = 10; Maximum epochs = 200; Activation = "ELU"; Optimizer = "RMSprop"; Loss = "MSE"; Batch size = 32
10	LGBM	Number of estimators = 1000; Learning rate = 0.01; Number of leaves = 35; Max depth = −1; Minimum child samples = 20; Minimum split gain = 0.001; Force col wise = True

Table 3 Performance of pricing model under different training-test ratio.

No.	Model	80–20%			70–30%			90–10%		
		MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE
1	MLR	19.5889	4.4259	2.0271	51.3050	7.1627	2.7931	10.3752	3.2211	1.7959
2	Lasso	4.6022	2.1453	1.7631	4.5476	2.1325	1.7439	4.2012	2.0497	1.6848
3	DT	3.1783	1.1199	1.7828	3.4990	1.8706	1.1850	3.1492	1.7746	1.0538
4	SVR	6.8124	2.6101	2.1994	6.7321	2.5946	2.1813	5.7918	2.4066	1.9890
5	MLP	1.7074	1.3067	0.9235	1.5713	1.2535	0.8629	0.9355	0.9672	0.7165
6	KNN	3.5501	1.8842	1.3464	3.5715	1.8898	1.3833	3.3319	1.8253	1.3472
7	GBDT	1.8630	1.3649	1.0129	1.8482	1.3595	1.0199	1.5875	1.2600	0.9567
8	LSTM	2.9336	1.7128	1.3407	4.1243	2.0308	1.6152	3.6693	1.9155	1.5013
9	RF	1.1682	1.0808	0.7661	1.3727	1.1716	0.8350	1.0718	1.0353	0.7284
10	LGBM	**0.9941**	**0.9970**	**0.6489**	**1.1497**	**1.0722**	**0.7266**	**0.8526**	**0.9234**	**0.6389**

The bold values indicate the best performance for each evaluation metric.

standard RBF kernel (C = 1.0, $\varepsilon = 0.1$), and LSTM is configured with 10 hidden units and the RMSprop optimizer. For LGBM, we conducted parameter tuning based on validation performance, selecting the configuration that minimized MSE on the development set (e.g., 1000 estimators, learning rate = 0.01). These standardized and optimized settings enhance both stability and comparability across baselines.

Evaluation metrics. To evaluate the accuracy of the pricing model relative to other benchmark models, three evaluation metrics were employed: mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) (Wang et al. 2023). The respective formulas are provided below:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{11}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{12}$$

Result

Performance of pricing model. To rigorously evaluate the performance of the proposed pricing model, a diverse set of models was utilized, including multiple linear regression (MLR), Lasso, decision tree, support vector machine (SVM), multilayer perceptron (MLP), K-nearest neighbors (KNN), gradient boosting decision tree (GBDT), long short-term memory network (LSTM), random forest (RF), and lightGBM. Experiments were conducted using different training-test ratios. The pricing error results for each model under various configurations are detailed in Table 3.

From the analysis of the data in Table 3, it can be observed that the MLR model exhibited the highest error rate across all three training-test ratio settings. In contrast, the LASSO model, which incorporates a regularization term for variable selection, demonstrated relatively lower errors. This finding highlights the limitations of linear pricing models when handling learning tasks with multiple features. For instance, under the 80–20 training-test ratio, the multilayer perceptron (MLP) was found to have the lowest pricing error among all machine learning models evaluated. Compared to support vector regression (SVR) and decision tree models, the MLP's use of nonlinear activation functions allows it to better capture and model complex nonlinear relationships within the data. Although traditional SVR and Decision Tree models can also address nonlinear problems, the MLP generally offers superior performance in uncovering deep, nonlinear relationships in the data. The Random Forest (RF), as a typical example of ensemble learning, also demonstrated good performance in the pricing task. Ultimately, the heterogeneous, data-driven LightGBM model constructed in our study showed the best pricing performance among all models. As illustrated in Fig. 6, it achieved the lowest prediction error across various training-test ratios, further confirming its exceptional performance.

To rigorously assess the statistical significance of predictive differences between models, we adopt the Diebold–Mariano (DM) test, a standard method for comparing the forecast accuracy of two models based on a loss differential series. The test evaluates the null hypothesis that the two models yield equal predictive performance, against the alternative that one model is significantly better.

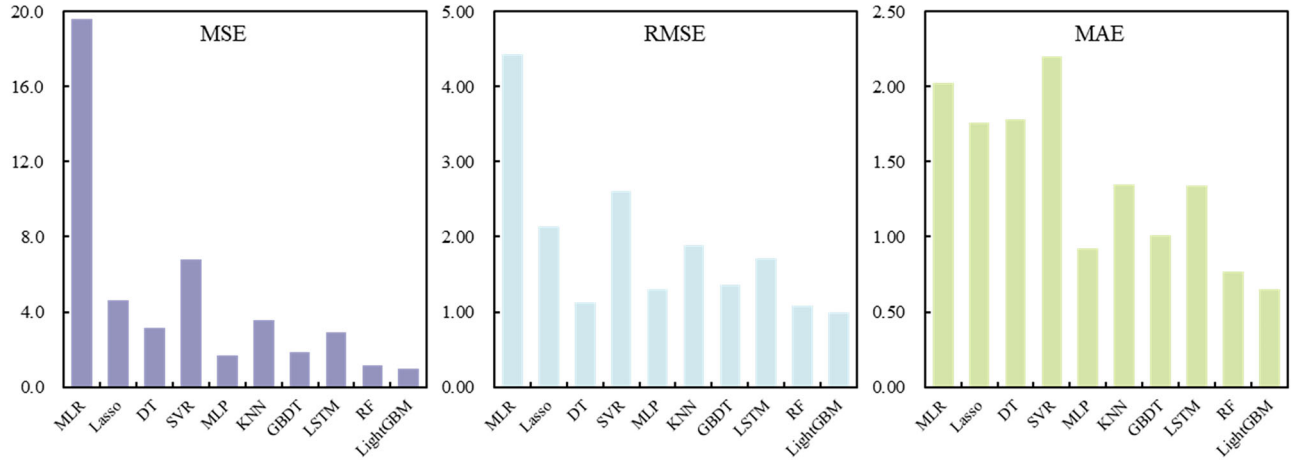

Fig. 6 Pricing error of models.

Table 4 DM statistical testing.

Model	Statistics	MLP	MLR	Lasso	DT	SVR	RF	GBDT	KNN	LSTM
MLR	DM	2.1468								
	p value	0.0327								
Lasso	DM	6.6999	1.7964							
	p value	0.0000	0.0735							
DT	DM	2.5278	1.9768	2.6904						
	p value	0.0120	0.0490	0.0076						
SVR	DM	9.8531	1.5245	6.6408	5.9888					
	p value	0.0000	0.1285	0.0000	0.0000					
RF	DM	1.8129	2.2082	11.7276	4.1610	13.9777				
	p value	0.0709	0.0280	0.0000	0.0000	0.0000				
GBDT	DM	0.5147	2.1259	9.2717	2.7080	11.8105	6.1548			
	p value	0.6072	0.0344	0.0000	0.0072	0.0000	0.0000			
KNN	DM	3.9512	1.9175	2.7687	0.6038	6.8364	6.4763	4.5025		
	p value	0.0001	0.0562	0.0060	0.5465	0.0000	0.0000	0.0000		
LSTM	DM	4.4679	1.9002	4.0846	1.0317	8.2878	9.3884	7.0814	0.4102	
	p value	0.0000	0.0584	0.0001	0.3031	0.0000	0.0000	0.0000	0.6820	
LGBM	DM	2.8854	2.2295	11.5785	4.4771	13.7249	1.9592	6.0502	6.9108	9.1490
	p value	0.0042	0.0266	0.0000	0.0000	0.0000	0.0511	0.0000	0.0000	0.0000

In this study, the DM test is conducted using the mean squared error (MSE) as the loss function and a forecast horizon of $h = 1$, which aligns with common practice in predictive regression. To account for finite sample effects, the Harvey-adjusted DM statistic is used, and statistical significance is assessed at the 5% level using a two-sided *t*-test.

All models are evaluated in pairwise comparisons, with the actual target values and predicted outputs used to compute the loss differentials. The DM statistics and corresponding *p* values are summarized in Table 4. As shown in the final row of Table 4, LGBM demonstrates statistically superior pricing performance over all baseline models, with most *p* values well below 0.05. For instance, when compared with MLP, the DM statistic is 2.8854 ($p = 0.0042$), allowing us to reject the null hypothesis and confirm the significance of the performance gain. This analysis supports the robustness and effectiveness of our proposed framework for data asset pricing. Therefore, the DM test results in the last row of Table 4 confirm that LGBM significantly outperforms all benchmark models, providing strong statistical evidence of its superior pricing performance.

Discussion. To validate the significance of alternative data in enhancing the accuracy of asset pricing, an 80–20 training-test ratio was employed to compare the pricing errors of models with and without the inclusion of non-traditional data, as shown in Table 5. Overall, most models exhibited a significant reduction in pricing error upon the incorporation of non-traditional data. For example, the LASSO model's mean squared error (MSE) decreased from 5.688 to 4.602 after including these data; correspondingly, the root mean squared error (RMSE) and mean absolute error (MAE) also showed reductions. Similar improvements were observed in models such as SVR, MLP, KNN, GBDT, LSTM, RF, and LGBM.

Notably, the proposed model in this study demonstrated a prediction error of 2.723 without the use of alternative data, which dropped to 0.994 after their inclusion, representing a 63.5% reduction in error. This indicates that non-traditional data, such as text, significantly enrich the data's dimensionality and quality by providing additional contextual information, capturing hidden features and patterns, enhancing the model's generalization ability, and improving fine-grained analysis, thereby substantially boosting prediction accuracy.

However, for the MLP model, the introduction of non-traditional data did not lead to a reduction in error. This can be attributed to two main reasons: (1) Transforming non-traditional data into features suitable for multivariate linear models may result in an extremely high-dimensional feature

Table 5 Pricing errors of models with and without alternative data.

NO.	Models	MSE			RMSE			MAE		
		Without	With	Δ	Without	With	Δ	Without	With	Δ
1	MLR	5.486	19.589	↑	2.342	4.426	↑	1.950	2.027	↑
2	Lasso	5.688	4.602	↓	2.385	2.145	↓	1.983	1.763	↓
3	DT	5.578	3.178	↓	2.362	1.120	↓	1.530	1.783	↑
4	SVR	7.273	6.812	↓	2.697	2.610	↓	2.267	2.199	↓
5	MLP	5.302	1.707	↓	1.787	1.307	↓	2.303	0.924	↓
6	KNN	6.223	3.550	↓	2.495	1.884	↓	1.997	1.346	↓
7	GBDT	2.815	1.863	↓	1.678	1.365	↓	1.292	1.013	↓
8	LSTM	3.510	2.934	↓	1.874	1.713	↓	1.455	1.341	↓
9	RF	2.419	1.168	↓	1.555	1.081	↓	1.143	0.766	↓
10	LGBM	2.723	0.994	↓	1.650	0.997	↓	1.246	0.649	↓

The symbol Δ represents the difference in pricing errors between using alternative data and not using alternative data. If the difference is greater than zero, it is indicated with an ↑, signifying that the alternative data has not reduced the pricing error. Conversely, if the difference is less than zero, it is indicated with a ↓, demonstrating that the alternative data has effectively reduced the pricing error.

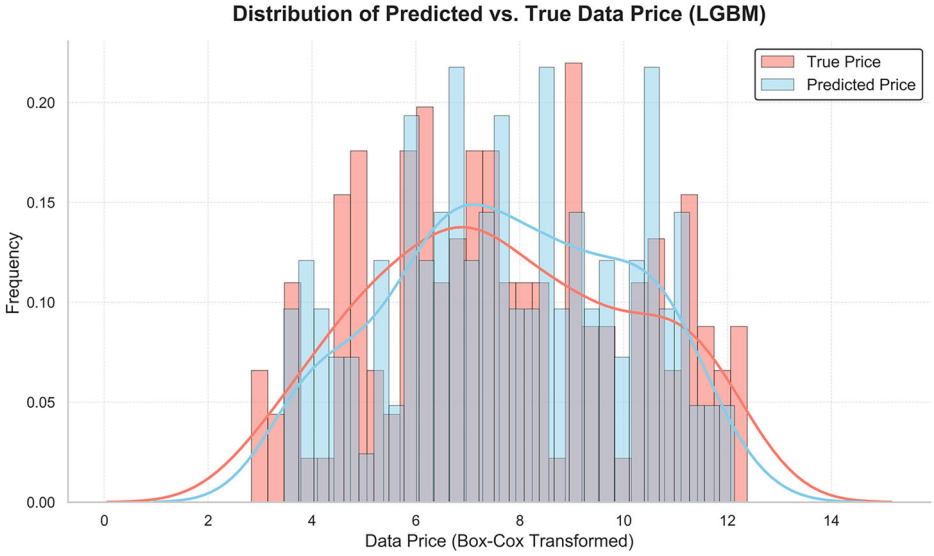

Fig. 7 Data price distribution comparison (LGBM).

space. In high-dimensional spaces, sample data becomes sparse, significantly increasing the difficulty of discovering meaningful patterns, leading to the so-called "curse of dimensionality"; (2) Alternative data often contain complex structures and relationships that may be nonlinear, whereas multivariate linear models are based on linear assumptions. Therefore, when the true relationships among data were nonlinear, linear models may fail to effectively capture these relationships, resulting in decreased prediction accuracy.

To evaluate the effectiveness of the model's estimation, Fig. 7 compares the distribution of the predicted data prices generated by the LGBM model with the true data prices (after Box-Cox transformation). The predicted price distribution (in blue) closely follows the true price distribution (in red), indicating that the model effectively captures the underlying data value structure. While slight deviations are observed in the upper and lower tails, the overall shape and central tendency of the two distributions are aligned. This result supports the fairness and robustness of the model's pricing mechanism, as it does not systematically overestimate or underestimate data price across the range. The consistency between predicted and true distributions provides further empirical justification for the reliability of the proposed multimodal valuation framework.

Scenario-based data value estimation

Evaluating text feature value through error reduction. Based on the comprehensive experimental results presented in the section "Result", it was observed that the integration of enhanced textual information significantly improves the predictive performance and accuracy of pricing models. In light of this finding, a more detailed strategy was adopted to investigate the specific value contribution and potential information gain of textual data in the pricing process. Textual information was decomposed into four core components: data asset titles, detailed data descriptions, target user groups, and detailed functional descriptions of the data.

In the experimental design for this phase, all traditional numerical features were kept constant to control for variables and isolate the impact of textual information on pricing performance. Subsequently, advanced BERT algorithms were employed to efficiently and accurately represent the semantics of the aforementioned four textual components. These semantic representations were then integrated as input features into the pricing model. The pricing performance under each input feature is shown in Table 6. This process aimed to systematically evaluate the specific effects of each type of textual information on enhancing the predictive capability of the pricing model through

Table 6 Pricing performances under different input features.

Methods	Features	80–20%			70–30%			90–10%		
		MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE
LGBM	TF	2.7226	1.6500	1.2457	2.6122	1.6162	1.2317	2.5761	1.6050	1.2520
	TF + Text (data title)	1.2715	1.1276	0.7684	1.5692	1.2527	0.8702	1.5022	1.2257	0.8198
	TF + Text (target user)	2.6050	1.6140	1.1421	2.4328	1.1199	1.5597	2.7867	1.6694	1.2409
	TF + Text (data function)	2.4041	1.5505	1.0854	2.2813	1.5104	1.0778	2.4970	1.5802	1.1401
	TF + Text (data descriptions)	**0.8016**	**0.8953**	**0.6248**	**0.9492**	**0.9743**	**0.6809**	**0.8116**	**0.9009**	**0.6173**
	TF + Text (total)	0.9941	0.9970	0.6489	1.1497	1.0722	0.7266	0.8526	0.9234	0.6389
MLP	TF	5.3025	1.7867	2.3027	5.4665	2.3381	1.8287	4.7131	2.1710	1.7092
	TF + Text (data title)	2.6124	1.6163	1.1653	2.8345	1.6836	1.1803	3.2863	1.8128	1.2544
	TF + Text (target user)	4.0210	2.0052	1.4659	3.8034	1.9502	1.4093	3.5362	1.8805	1.3539
	TF + Text (data function)	4.4767	2.1158	1.6672	4.5859	2.1415	1.6115	3.2846	1.8124	1.3204
	TF + Text (data descriptions)	**1.0651**	**1.0320**	**0.6483**	**1.1049**	**1.0511**	**0.7034**	**0.9007**	**0.9491**	**0.6676**
	TF + Text (total)	1.7074	1.3067	0.9235	1.5713	1.2535	0.8629	0.9355	0.9672	0.7165
DT	TF	5.5779	2.3618	1.5295	5.2832	2.2985	1.5048	5.5555	2.3570	1.5112
	TF + Text (data title)	4.8629	2.2052	1.4432	5.0400	2.2450	1.5140	4.2060	2.0509	1.2653
	TF + Text (target user)	5.6352	2.3739	1.5087	4.5270	2.1277	1.3550	5.7507	2.3981	1.5182
	TF + Text (data function)	4.1051	2.0261	1.2680	4.7251	2.1737	1.3668	5.2672	2.2950	1.5241
	TF + Text (data descriptions)	**2.9486**	**1.7171**	**1.0972**	**3.1107**	**1.7637**	**1.1673**	**3.1418**	**1.7725**	**1.1185**
	TF + Text (total)	3.1783	1.1199	1.7828	3.4990	1.8706	1.1850	3.1492	1.7746	1.0538
GBDT	TF	2.8153	1.6779	1.2922	2.8486	1.6878	1.2866	2.6176	1.6179	1.3153
	TF + Text (data title)	1.8530	1.3612	1.0152	2.0159	1.4198	1.0273	1.8415	1.3570	0.9990
	TF + Text (target user)	2.8905	1.7002	1.2572	2.7421	1.6559	1.2455	2.7182	1.6487	1.2887
	TF + Text (data function)	2.8044	1.6746	1.2172	2.4539	1.5665	1.1528	2.3408	1.5300	1.1697
	TF + Text (data descriptions)	**1.1892**	**1.0905**	**0.8086**	**1.3966**	**1.1818**	**0.8818**	**1.2838**	**1.1331**	**0.8391**
	TF + Text (total)	1.8630	1.3649	1.0129	1.8482	1.3595	1.0199	1.5875	1.2600	0.9567
RF	TF	2.4185	1.5552	1.1429	2.4799	1.5748	1.1575	2.6446	1.6262	1.2125
	TF + Text (data title)	1.7186	1.3110	0.9076	1.9694	1.4033	0.9913	1.7925	1.3388	0.9372
	TF + Text (target user)	2.5307	1.5908	1.1290	2.3857	1.5446	1.0973	2.8026	1.6741	1.1928
	TF + Text (data function)	2.3676	1.5387	1.0450	2.3092	1.5196	1.0264	2.5204	1.5876	1.0876
	TF + Text (data descriptions)	**1.1402**	**1.0678**	**0.7340**	**1.2443**	**1.1155**	**0.8120**	**1.0117**	**1.0058**	**0.6861**
	TF + Text (total)	1.1682	1.0808	0.7661	1.3727	1.1716	0.8350	1.0718	1.0353	0.7284

TF means traditional features.
The bold values indicate the best performance for each evaluation metric.

comparative experiments. The goal was to provide a robust empirical foundation and theoretical basis for optimizing data pricing strategies.

First, through a comparative analysis of the application effects of traditional numerical features and textual features in pricing models, the following conclusions were drawn: the introduction of textual information as a supplement to traditional numerical features significantly enhances the pricing performance of the model. For instance, using the light gradient boosting machine (LGBM) model under an 80:20 train-test split, the mean squared error (MSE) reached 2.7226 when relying solely on traditional features (assumed to be a certain set of numerical features, denoted as TF). This indicates a non-negligible level of prediction bias. Subsequently, four types of textual information—data descriptions, data titles, target groups, and data function descriptions—were individually incorporated into the pricing model. The experimental results demonstrated that the inclusion of these textual features reduced the MSE to varying extents, thereby improving pricing accuracy. Notably, the addition of data descriptions resulted in the most significant reduction in error, with the MSE decreasing to 0.8016, highlighting the critical role of data descriptions in enhancing pricing precision. Furthermore, the inclusion of data titles also markedly reduced the MSE to 1.2715, indicating that data titles contain substantial pricing-related information. In contrast, while the addition of target group and data function descriptions also led to a reduction in error, their impact was comparatively limited relative to data descriptions and data titles. This observation was consistently validated in experiments with two other different train-test split ratios, further reinforcing our conclusion: textual information, particularly data descriptions and data titles, should be considered key factors in improving the performance of pricing models.

Second, through comparative experiments analyzing the application effects of all textual information versus four subcategories of textual information (i.e., data titles, target groups, data function descriptions, and data descriptions) in pricing models, several important findings were obtained. When all textual information was used as input, the model captured a broader range of data value information. However, the pricing error was lower when only data titles, target groups, and data function descriptions were included, yet still higher than when the best subcategory of textual information (i.e., data descriptions) was used alone. This comparison reveals the dual characteristics of textual information in data pricing: on one hand, it indeed contains rich data value information that enhances the accuracy of pricing models; on the other hand, it inevitably includes a certain degree of redundant information, which can interfere with the model's predictive capability. Specifically, information such as target users, data function descriptions, and data titles, when used individually or in combination, did not provide the same degree of pricing optimization as data descriptions. Instead, they may have introduced unnecessary complexity and noise, leading to information redundancy. Therefore, the findings of this study emphasize the need for careful selection and optimization of textual information when constructing pricing models. This approach aims to maximize the value of textual information while minimizing

the negative impact of redundant information on model performance.

Third, to verify whether the role of textual description information in enhancing data pricing performance is solely attributable to the uniqueness of a specific method, a selection of representative machine learning algorithms was made based on the experimental results in the section "Performance of pricing model". These algorithms include multilayer perceptron (MLP), decision tree (DT), gradient boosting decision tree (GBDT), and random forest (RF). The objective was to comprehensively examine the performance differences of various textual information across different pricing methods. Through a series of comparative experiments, the impact of different types of textual information on data pricing accuracy was systematically analyzed. The experimental results consistently demonstrated that when utilizing the four pricing methods, the inclusion of data descriptions significantly reduced pricing errors. Specifically, the introduction of data descriptions led to notable improvements in key error metrics, including mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). This further corroborates the general effectiveness and importance of textual description features in enhancing the accuracy of pricing models.

In summary, this subsection systematically evaluates the impact of different text features on data pricing performance. Results show that integrating textual features—especially data descriptions and titles—significantly enhances model accuracy across various algorithms. However, not all textual inputs are equally beneficial. Elements such as target users and functional descriptions may introduce redundancy. These findings highlight the importance of selective text integration in multimodal pricing and offer practical insights for optimizing textual inputs in real-world valuation tasks.

Measuring data value through feature exclusion. To quantitatively assess the specific value or contribution of different types of features within the pricing mechanism, this study proposes an innovative use of SHAP value theory as an analytical tool. Given that textual data, when processed through the BERT model, is transformed into a high-dimensional (768-dimensional) information vector, it becomes challenging to directly integrate it with traditional numerical features within the same value evaluation framework. To address this technical bottleneck, a supervised learning model based on the multilayer perceptron (MLP) was meticulously designed. This model aims to effectively reduce the dimensionality of the high-dimensional textual representations to a one-dimensional feature space, thereby enabling compatibility and unified analysis with numerical features.

Specifically, the MLP model was utilized to successfully map four key dimensions of textual information—data asset titles, detailed data descriptions, target user demographics, and detailed functional descriptions—each into a single-dimensional numerical feature. This process not only preserved the essential value components of the textual information but also significantly simplified the subsequent feature value estimation process. This laid a solid foundation for comprehensively evaluating the multidimensional contributions of data assets in the pricing context. Through this innovative approach, the study effectively overcame the analytical challenges posed by the high-dimensional representation of textual data, providing new perspectives and tools for deep exploration in the field of data pricing.

Subsequently, the SHAP values for four textual features and twelve traditional numerical features were systematically calculated and graphically presented in Fig. 8. Figure 8 clearly and intuitively reveals the relative importance of each feature in the data pricing process. Specifically, the data description emerged as the core feature, exerting significantly greater influence than all other considered factors, particularly surpassing the next most important feature, the data title. Moreover, a detailed analysis of Fig. 8 indicates that numerical features such as data size, usage frequency, data type, and data value rating also play crucial roles in the data pricing mechanism. These features collectively form the multidimensional framework for assessing data value and determining its market price. This finding not only deepens our understanding of the complexities involved in data pricing but also provides robust data support and a theoretical basis for subsequent optimization of data asset management and trading strategies.

Upon delving into the importance of features, a rigorous ranking of features based on SHAP values was conducted. This step aimed to systematically evaluate the contribution of each data feature within the pricing model. Subsequently, a strategic feature selection process was implemented, wherein features of extreme importance (both high and low value) were incrementally removed. The impact of these adjustments on pricing performance was empirically tested by retraining the model. To objectively and comprehensively assess the effectiveness of this strategy, three sets of experiments were meticulously designed, each based on different training-to-testing ratios to simulate various data usage scenarios. Three error metrics—MSE, MAE, and RMSE—were selected to quantify and compare the predictive accuracy of the model under different experimental conditions. As illustrated in Fig. 9, these error metrics were graphically presented to show their variations across different training-to-testing ratios. This visualization aimed to provide an in-depth analysis of the specific impact and potential patterns resulting from the removal of high- and low-value features on the performance of the pricing model.

Based on the analysis of the experimental results, several important conclusions can be drawn:

Firstly, high-value data features occupy an indispensable core position within the model. Specifically, when these key features are incrementally removed, the pricing performance of the model exhibits a significant decline, characterized by a sharp increase in pricing errors, followed by a relatively stable fluctuation phase. This phenomenon underscores the critical informational value of the removed features and their direct contribution to the model's predictive accuracy. For instance, as shown in Fig. 9A, under the 70–30% training-to-testing ratio configuration, the MSE, MAE, and RMSE error metrics all experienced a rapid increase as the number of high-value features removed increased. When the removal reached three features, the error growth trend temporarily stabilized until six features were removed; however, when the removal reached seven features, the model's performance significantly deteriorated, with pricing errors sharply escalating. Notably, the MSE error surged to 8.2, compared to 1.18 when no features were removed, representing an increase of nearly 5.9-fold. This stark contrast highlights the crucial importance of high-value features in maintaining the model's predictive accuracy. Additionally, it is worth noting that this phenomenon was also validated in experiments with the other two different training-to-testing ratio settings, further reinforcing the irreplaceable role of high-value data features in the pricing model.

Secondly, the impact of low-value data features on the model's pricing performance exhibits significant bidirectional effects. The data presented in Fig. 8 clearly reveals this phenomenon: during the removal of low-value features, the model's pricing performance undergoes complex changes. Specifically, the initial removal of some low-value features results in a notable reduction in prediction errors, indicating that unfiltered low-value information may negatively affect the model's performance by

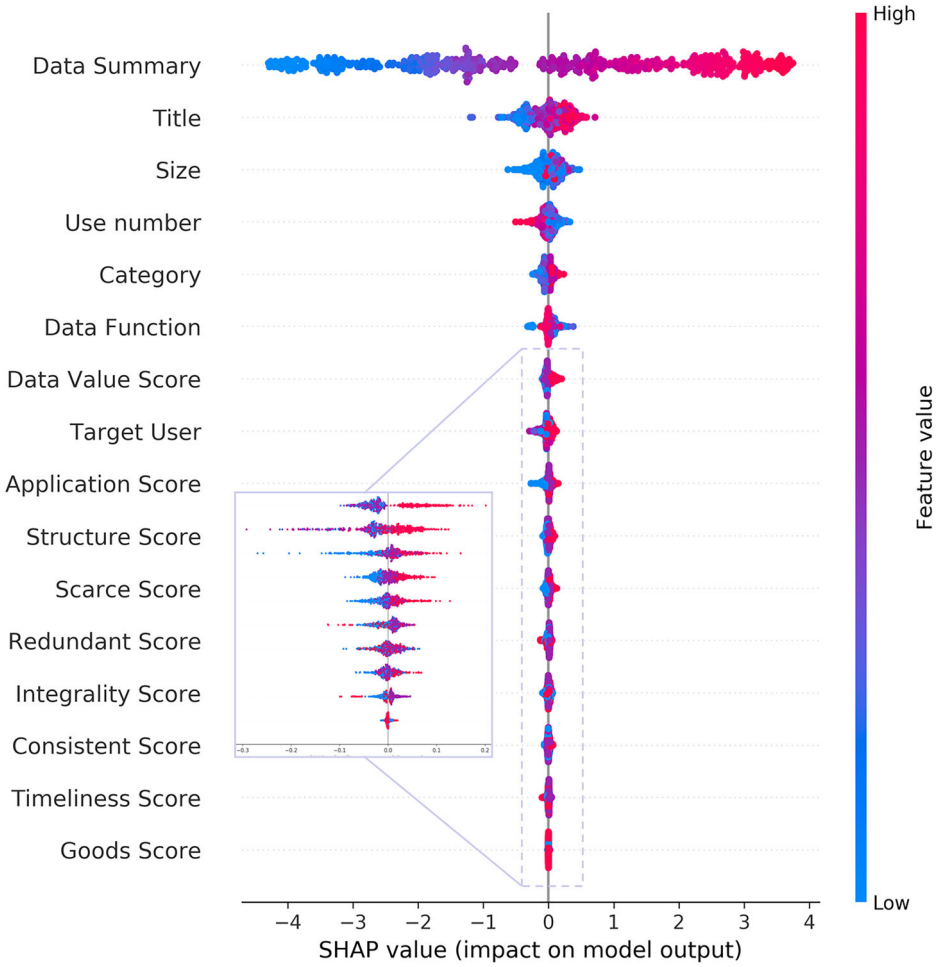

Fig. 8 SHAP value.

introducing noise or redundant information that interferes with accurate predictions. However, as the removal process continues, a contrasting trend is observed—pricing errors begin to increase. This pivotal change suggests that even within data deemed as low-value, there exist elements that contribute positively to pricing decisions. Their removal, therefore, leads to a decline in the model's performance.

In conclusion, high-value data features are undoubtedly critical for enhancing the performance of pricing models, as they positively contribute to the model's accuracy and reliability. In contrast, low-value data features exhibit a complex dual role. They can act as sources of noise, adversely affecting the model, or harbor useful signals essential for the pricing process. This necessitates the adoption of more refined feature selection strategies during model construction and optimization to fully balance and maximize the contributions of various data features to the model's performance.

In summary, this subsection highlights that while high-value features are essential for accurate pricing, selectively removing low-value features can initially improve performance by reducing noise. However, excessive exclusion may discard useful signals. These findings highlight the importance of refined feature selection in optimizing model robustness and interpretability.

Conclusion

This study proposes an innovative data asset pricing framework based on alternative data. This framework integrates textual information related to data functions with standard numerical characteristics, significantly enhancing the model's pricing capabilities. Empirical research demonstrates that the developed pricing model exhibits superior predictive performance compared to traditional machine learning and deep learning models. Non-traditional data centered around textual information significantly enriches contextual content, reveal latent features and patterns, enhances the model's generalization ability, and increases the precision of detailed analysis, thereby significantly boosting the dimensionality and quality of the data and, consequently, the accuracy of predictions. Additionally, a scenario-based method for measuring data value has been established. By utilizing multilayer perceptrons to map textual information, the challenges posed by the high-dimensional representation of text data are effectively overcome. Furthermore, SHAP is introduced for data feature ranking, and features of extreme importance (both high-value and low-value) are progressively removed to examine their specific impacts on pricing performance. High-value data features are undoubtedly critical for enhancing the performance of pricing models. In contrast, low-value data features exhibit a complex dual role.

Our research not only enhances the accuracy of data asset pricing but also deepens the understanding of the application value of non-traditional data in pricing models, which is essential for data pricing and management. Furthermore, this study paves the way for future explorations, such as incorporating a wider variety of data sources (e.g., images, audio, and video data) and employing more advanced deep learning and ensemble learning techniques to further improve the performance of pricing models.

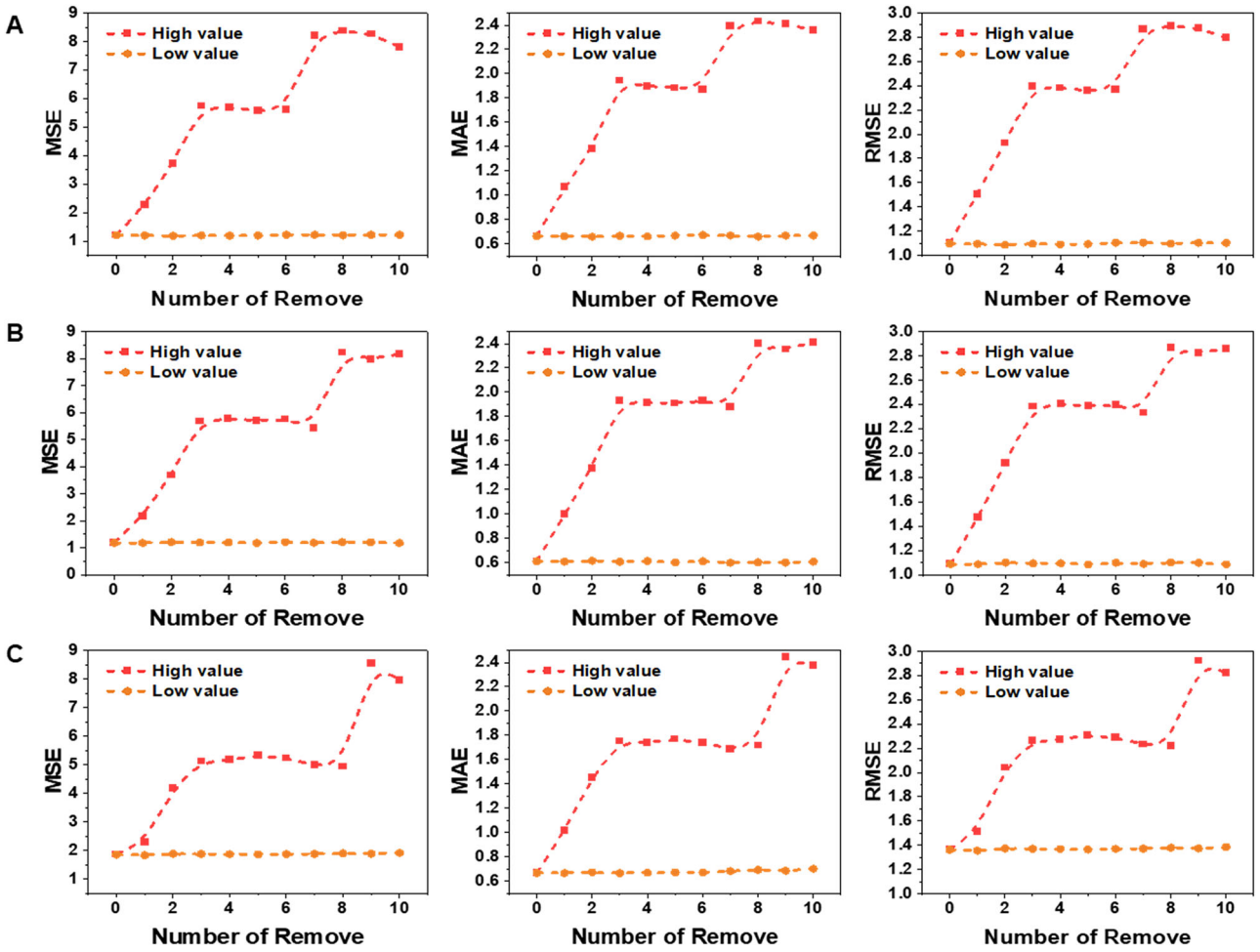

Fig. 9 Error trends after feature value removal. The trends observed after removing high- and low- value features under different training-to-testing ratios:(A) 70-30%, (B) 80-20%, and (C) 90-10%.

At the same time, we recognize that the reliability and accuracy of functional textual descriptions play a crucial role in determining data value. In practical settings, such descriptions may suffer from inconsistency, ambiguity, or even intentional manipulation. To address this, future work could incorporate automatic quality screening techniques (e.g., coherence or factuality scoring), cross-verification from multiple information sources, or the integration of domain-specific knowledge bases to enhance textual credibility. Such refinements would further strengthen the robustness and fairness of the pricing model.

Data availability

The datasets collected and analyzed during the current study should be retrieved upon reasonable requests from the corresponding author based on appropriate reasons.

Received: 24 April 2025; Accepted: 18 September 2025;
Published online: 17 November 2025

References

Abbasi M, Prieto J, Shahraki A, Corchado JM (2023) Industrial data monetization: a blockchain-based industrial IoT data trading system. Internet Things 24:100959

Alonso-Robisco A, Carbo JM (2023) Analysis of CBDC narrative by central banks using large language models. Financ Res Lett 58:104643

Anand R, Jha JK (2022) Multi-period dynamic pricing model for deteriorating products in a supply chain with preservation technology investment and carbon emission. Comput Ind Eng 174:108817

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc Ser B-Stat Methodol 26211–252

Cao X, Chen Y, Liu KJR (2017) Data trading with multiple owners, collectors, and users: an iterative auction mechanism. IEEE Trans Signal Inf Process Netw 3:268–281

Dessaint O, Foucault T, Fresard L (2024) Does alternative data improve financial forecasting? The horizon effect. J Financ 79:2237–2287

Djeundje VB, Crook J, Calabrese R, Hamid M (2021) Enhancing credit scoring with alternative data. Expert Syst Appl 163:113766

Feng Z, Yu S, Zhu Y (2023) Towards personalized privacy preference aware data trading: a contract theory based approach. Comput Netw 224:109637

Gao J, Li P, Chen ZK, Zhang JN (2020) A survey on deep learning for multimodal data fusion. Neural Comput 32:829–864

Gneezy A, Gneezy U, Lauga DO (2014) A reference-dependent model of the price-quality heuristic. J Mark Res 51:153–164

Gu SH, Kelly B, Xiu DC (2021) Autoencoder asset pricing models. J Econ 222:429–450

Hansen KB, Borch C (2022) Alternative data and sentiment analysis: prospecting non-standard data in machine learning-driven finance. Big Data Soc 9:20539517211070701

Hao J, Deng Z, Li J (2023a) The evolution of data pricing: from economics to computational intelligence. Heliyon 9:e20274

Hao J, Feng QQ, Li JP, Sun XL (2023) A bi-level ensemble learning approach to complex time series forecasting: taking exchange rates as an example. J Forecast 42:1385–1406

Hao J, Yuan J, Li J (2024) HCEG: a heterogeneous clustering ensemble learning approach with gravity-based strategy for data assets intelligent pricing. Inf Sci 678:121082

Hao J, Yuan J, Li J, Liu M, Liu Y (2023b) Ensemble pricing model for data assets with ranking-pruning-averaging strategy. Procedia Comput Sci 221:813–820
Henry I, Abiola A, Eseosa O (2023) Competitive behaviour of major GSM firms' internet data pricing in Nigeria: a game theoretic model approach. Heliyon 9:e12886
Hlongwane R, Ramaboa KKKM, Mongwe W (2024) Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data. PLos ONE 19:e0303566
Jiang P, Liu Z, Abedin MZ, Wang J, Yang W, Dong Q (2024) Profit-driven weighted classifier with interpretable ability for customer churn prediction. Omega 125:103034
Kikuchi S, Kronprasert N (2012) Effects of data quality and quantity in systems modelling: a case study. Int J Gen Syst 41:697–711
Koutris P, Upadhyaya P, Balazinska M, Howe B, Suciu D (2015) Query-based data pricing. J ACM 62:43
Lehrer S, Xie T, Zeng T (2021) Does high-frequency social media data improve forecasts of low-frequency consumer confidence measures? J Financial Econ 19:910–933
Li Y, Chao X, Ercisli S (2022) Disturbed-entropy: a simple data quality assessment approach. ICT Express 8(3):309–312
Liang J, Yuan CH (2021) Data price determinants based on a hedonic pricing model. Big Data Res 25:100249
Liu K, Qiu X, Chen W, Chen X, Zheng Z (2019) Optimal pricing mechanism for data market in blockchain-enhanced internet of things. IEEE Internet Things J 6:9748–9761
Liu SQ, Wang JM, Li Q (2023) Alternative data and trade credit financing: evidence from third-party online sales disclosure. Financ Res Lett 58:104469
Ma X, Che T, Jiang Q (2025) A three-stage prediction model for firm default risk: an integration of text sentiment analysis. Omega 131:103207
Mehta S, Dawande M, Janakiraman G, Mookerjee V (2021) How to sell a data set? Pricing policies for data monetization. Inf Syst Res 32:1281–1297
Miao X, Gao Y, Chen L, Peng H, Yin J, Li Q (2022) Towards query pricing on incomplete data. IEEE Trans Knowl Data Eng 34:4024–4036
Mumbower S, Garrow LA (2014) Data set online pricing data for multiple US carriers. Manuf Serv Oper Manag 16:198–203
Nguyen Cong L, Hoang DT, Wang P, Niyato D, Kim DI, Han Z (2016) Data collection and wireless communication in internet of things (IoT) using economic analysis and pricing models: a survey. IEEE Commun Surv Tutor 18:2546–2590
Niu C, Zheng Z, Wu F, Tang S, Chen G (2022) Online pricing with reserve price constraint for personal data markets. IEEE Trans Knowl Data Eng 34:1928–1943
Peyvandi A, Majidi B, Peyvandi S, Patra JC (2022) Privacy-preserving federated learning for scalable and high data quality computational-intelligence-as-a-service in Society 5.0. Multimed Tools Appl 81:25029–25050
Sheng Y, Qu Y, Ma D (2024) Stock price crash prediction based on multimodal data machine learning models. Financ Res Lett 62:105195
Tian Y, Ding Y, Fu S, Liu D (2022) Data boundary and data pricing based on the Shapley value. IEEE Access 10:14288–14300
Tigges M, Mestwerdt S, Tschirner S, Mauer R (2024) Who gets the money? A qualitative analysis of fintech lending and credit scoring through the adoption of AI and alternative data. Technol Forecast Soc Change 205:123491
Veldkamp L (2023) Valuing data as an asset. Rev Financ 27:1545–1562
Wang J, Zhuang Z, Gao D (2023) An enhanced hybrid model based on multiple influencing factors and divide-conquer strategy for carbon price prediction. Omega-Int J Manag Sci 120:102922
Xiao Z, He D, Du J (2021) A Stackelberg game pricing through balancing trilateral profits in big data market. IEEE Internet Things J 8:12658–12668
Xu J, Hong N, Xu Z, Zhao Z, Wu C, Kuang K, Wang J, Zhu M, Zhou J, Ren K, Yang X, Lu C, Pei J, Shum H (2022) Data-driven learning for dta rights, data pricing, and privacy computing. Engineering 25:66–76
Xu L, Jiang C, Qian Y, Zhao Y, Li J, Ren Y (2017) Dynamic privacy pricing: a multi-armed bandit approach with time-variant rewards. IEEE Trans Inf Foren Sec 12:271–285
Xu W, Jiang L, Li C (2021) Improving data and model quality in crowdsourcing using cross-entropy-based noise correction. Inf Sci 546:803–814
Yang J, Zhao C, Xing C (2019) Big data market optimization pricing model based on data quality. Complexity. https://doi.org/10.1155/2019/5964068
Yu H, Zhang M (2017) Data pricing strategy based on data quality. Comput Ind Eng 112:1–10
Yu M, Wang J, Yan J, Chen L, Yu Y, Li G, Zhou M (2022) Pricing information in smart grids: a quality-based data valuation paradigm. IEEE Trans Smart Grid 13:3735–3747
Yuan J, Li J, Hao J (2023) A dynamic clustering ensemble learning approach for crude oil price forecasting. Eng Appl Artif Intell 123:106408
Zhang M, Beltran F, Liu J (2023) A survey of data pricing for data marketplaces. IEEE Trans Big Data 9:1038–1056
Zhang X, Xia Z, He F, Hao J (2024) Forecasting crude oil prices with alternative data and a deep learning approach. Ann Oper Res. https://doi.org/10.1007/s10479-024-06056-8
Zhang X, Yue WT, Yu Y, Zhang X (2023) How to monetize data: an economic analysis of data monetization strategies under competition. Decis Support Syst 173:114012
Zhang Z, Liu G, Wu J, Tan Y (2022) Data and algorithm pricing: incentive mechanisms design for federated learning. SSRN https://ssrn.com/abstract=4061980

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (T2293774, 72571269 and 72201265), National Key Research and Development Program of China (2022YFC3321104), and China Postdoctoral Science Foundation-funded project (2023T160635 and 2022M723105).

Author contributions

JH was responsible for conceptualization, investigation, methodology development, software implementation, and manuscript drafting. ZD contributed to experimental analysis, validation, software coding, and data curation. JL (Jin) participated in manuscript revision, editing, and funding acquisition. JL (Jianping) supervised the project, managed administration, secured funding, and oversaw manuscript revision. All authors critically reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Jun Hao or Jianping Li.