ISSN 2616-5902 Vol. 6, Issue 11: 42-46, DOI: 10.25236/AJBM.2024.061107

Research on Data Asset Feature Analysis and Value Evaluation Strategy Based on Random Forest and B-P Neural Network

Yijiao Fan[*]

JP Morgan, New York, 11101, USA
yijiaofanedit@outlook.com
[*]*Corresponding author*

Abstract: *Against the backdrop of the booming development of the digital economy, this article delves into the importance of evaluating the value of data assets, pointing out that as the digital economy becomes a new driving force for national economic development, the status of data as a production factor is highlighted and transformed into an important asset. However, the current accounting system's inclusion of data assets is not yet complete, which limits the accurate measurement of data value and its economic contribution, and affects the healthy development of the data trading market. The theoretical contribution of this article lies in the comprehensive analysis of the influencing factors of data asset value using B-P neural network and random forest algorithm, and the construction of a high prediction accuracy evaluation model. It was found that data capacity, size, quality, freshness, and industry all significantly affect its value, with the random forest model being preferred due to its advantages. This study not only enriches the theoretical system but also provides new ideas for practical applications. The established model can guide data asset evaluation, support market fair trading and resource optimization allocation, and propose policy recommendations to promote data quality improvement, market perfection, and value maximization. However, research also faces challenges such as single data sources and limited data types. In the future, it is necessary to expand data sources, explore nontextual numerical data evaluation methods, and introduce more advanced technologies to more comprehensively and accurately evaluate the value of data assets, promote the improvement and development of the data factor market, and assist in the digital transformation of the economy.*

Keywords: *Valuation of Data Assets, Digital Economy, Machine Learning Algorithms, Data Trading Markets, Economic Digital Transformation*

1. Introduction

In the booming wave of the digital economy, data has been elevated to a key production factor on par with capital, land, labor, and technology, and has transformed into an indispensable asset. This transformation not only highlights the importance of data in terms of information and knowledge attributes but also profoundly affects every corner of the economy and society. From Laney's 3Vs model to expanding to the 5Vs model, big data continues to enrich our understanding of the essence of data with its diverse features such as capacity, speed, diversity, accuracy, and value. As a fusion of experimental data and social data, big data is not only a product of technological progress but also the core driving force of the secondary economy. However, the current accounting system's neglect of data assets makes it difficult to quantify the value of data and its contribution to economic growth, hindering the prosperity of the data trading market and the efficient allocation of resources. To solve this problem, this article focuses on using B-P neural networks and random forest algorithms in machine learning to deeply analyze the multidimensional influencing factors of data asset value, and construct a high-precision evaluation model, aiming to open up new paths for the theoretical exploration and practical application of data asset value. This study not only enriches the theoretical framework of data asset evaluation but also provides strong support for promoting data-fair trade and accelerating the healthy development of the digital economy. Data assets are gradually becoming a new engine for promoting economic and social development by reducing information asymmetry, lowering risks, and synergizing with other production factors. Therefore, the valuation and pricing of data assets need to comprehensively consider their multiple attributes such as technology, economy, and society, to fully reflect their core value in the digital economy era.

2. Correlation Theory

2.1. Definition of Concept

In the vibrant and ever-evolving digital realm, unlocking the boundless treasures within data assets [1]—those elusive yet invaluable virtual jewels—demands a deeply personalized and nuanced valuation approach. This intricate journey delves into the data's lifecycle, meticulously accounting for expenses from arduous mining to meticulous cleansing, intricate processing, artful structuring, and seamless dissemination. Beyond tangible costs, we embrace the whispers of indirect expenses and hardware fatigue, enriching the story of data transformation. At its heart, we cherish the essence of quality, transcending metrics to embrace accessibility, confidentiality, portability, and traceability. The freshness of timely updates invigorates relevance, while capacity, grandeur of size, the elegance of structuring, and purity of cleanliness conspire to shape practicality and growth potential. This vibrant and adaptable framework resonates with the human spirit, celebrating data's invaluable role in shaping our interconnected future.

2.2. Theoretical Basis

The market value method is limited by the imperfect trading market and the personalized characteristics of data assets, resulting in a scarcity of comparable objects; The reset cost method makes it difficult to accurately measure the physical and functional depreciation of data assets due to their intangibility; Although the present value of earnings method is theoretically applicable, it faces challenges in practical operation such as the nonmateriality of data, complex interweaving of earnings, and uncertainty of future earnings. With the increasing frequency of data trading and the urgent need for accurate pricing, although traditional methods and some improved methods, such as user perceived value based methods, attempt to combine market methods with subjective feedback, they are questioned due to market limitations, strong subjectivity, and lack of objectivity. In this context, scholars have begun to explore the application of machine learning in data asset evaluation. By utilizing the self-learning and adjustment capabilities of machine learning, as well as the multi-threaded ability to handle complex nonlinear problems, and combining the advantages of big data computing and complex scene recognition, a new path has been opened up for the value evaluation of data assets. This article aims to construct an accurate evaluation model by applying machine learning methods to overcome the limitations of traditional methods and achieve a scientific and objective assessment of the value of data assets.

3. Research Method

3.1. Machine Learning

Machine learning is one of the core areas of artificial intelligence [2], focusing on automatically processing data, with the main goal of using these patterns to predict or make decisions on unknown data. With the continuous advancement of technology, the application of machine learning has gradually expanded to multiple fields. Although there is no unified definition, its core concept has been widely applied. Machine learning is usually divided into three types: supervised learning, unsupervised learning, and reinforcement learning, each with its unique application scenarios. In supervised learning, the model improves performance by adjusting the predicted results to match known reference results; Unsupervised learning identifies potential trends by analyzing the intrinsic features of data; Semi-supervised learning combines the characteristics of supervised learning and unsupervised learning, and is used to process partially labeled and unlabeled data. The basic principle of machine learning is to extract features from a large amount of historical data and build complex adaptive models through training, enabling them to solve specific problems or make predictions.

3.2. B-P Neural Network

B-P neural network is a feedforward multi-layer perceptron network widely used in the field of supervised learning. Its design inspiration comes from the operation of biological neurons, which can effectively handle systems with nonlinearity, multiple factors, uncertainty, randomness, and dynamic changes. Through the mechanism of error backpropagation, B-P neural networks have successfully modeled and predicted many complex problems. The intricate architecture of Backpropagation (B-P) neural networks, comprising an input layer, multiple hidden layers, and an output layer interconnected by weighted connections, facilitates precise data manipulation through forward propagation, where input signals are processed using weighted sums and Tanh activation functions [3]. Should the output error exceed a preset threshold, the network seamlessly transitions to backpropagation, iteratively adjusting weights and thresholds via gradient descent to minimize error and refine accuracy. This adaptability and precision underscore B-P neural networks' significant potential in tasks like data asset valuation, offering unparalleled insights into assessing the true worth of assets in today's data-driven economy.

3.3. Random Forest

Random Forest [4] is an advanced combination classifier algorithm proposed by Leo Breiman in 2001, which combines Bagging ensemble learning theory and random subspace methods to become a powerful nonlinear modeling tool. This method randomly extracts multiple sub-sample sets from the original dataset using the Bootstrap resampling technique [5], with each sub-sample set having the same size as the original dataset. Then, based on each sub-sample set, a decision tree is constructed as the base learner, and finally, the prediction results of all decision trees are fused through majority voting (for classification problems) or averaging (for regression problems) to generate the final prediction. Random forests significantly improve the accuracy of predictions, while exhibiting good robustness to outliers and noisy data, and effectively reducing the risk of overfitting. This makes random forests particularly suitable for dealing with problems that lack empirical references, unclear rules, multiple constraints, and missing data. It is easy to operate and computationally efficient, overcoming the shortcomings of traditional prediction methods in terms of insufficient information acquisition and low efficiency, and providing a reliable prediction foundation for practical applications. The construction of random forests relies on decision trees and Bagging ensemble algorithms. As the basic unit of a random forest, the decision tree constructs the model by continuously dividing nodes, and its core is to select the optimal partition based on features. In regression problems, such as predicting the value of data assets, decision trees use the square error minimization criterion for feature selection. The Bagging algorithm [6] extracts multiple sub-sample sets from the original dataset through random sampling and independently trains a decision tree for each sub-sample set. Finally, the prediction results of all decision trees are averaged to obtain the final prediction value. In addition, the "out of bag data" in the Bagging algorithm can be used to evaluate the generalization error of the model without the need for an additional test set. By combining the advantages of decision trees and Bagging algorithms, random forests achieve efficient and accurate predictive performance.

4. Results and Discussion

4.1. Data Acquisition and Preprocessing Stage

In the data acquisition and preprocessing stage, this article confirmed that the data source was Youyi Data Network, and successfully collected 5822 data asset transaction samples across time and industries through Python crawler technology. After deep cleaning and preprocessing, including text mining to extract key information, deleting invalid and duplicate data, identifying and removing outliers [7] and outliers, and handling missing values, 3996 valid samples were ultimately retained. In terms of feature selection and variable definition, multiple evaluation indicators including data capacity, data size, freshness, industry classification, etc. were determined based on domestic and foreign research results and data asset characteristics, and the entropy weight method was used to calculate the comprehensive score of data quality. By constructing a decision matrix and normalizing it, this article ensures a fair comparison of all evaluation indicators on a unified scale and then calculates information entropy and entropy weight to reflect the importance of each indicator, ultimately obtaining the comprehensive score of data quality for each data asset sample. This process comprehensively considers the quantitative and qualitative characteristics of the data, ensuring the objectivity and comprehensiveness of the evaluation, displaying the definitions and descriptive statistical results of each variable, highlighting the important impact of data capacity on the value of data assets, and the necessity of using machine learning evaluation, providing reliable data support for subsequent model construction.

4.2. Implementation of Data Asset Value Evaluation Model

When delving into the data asset valuation model based on the B-P (backpropagation) neural network, we utilized the neural net package in R language for systematic model construction and parameter tuning. Through meticulous empirical analysis, we discovered that a single hidden layer neural network, with its nodes, gradually increased from 3 to 6, exhibited good prediction accuracy but encountered non-convergence with excessive nodes, emphasizing the need for balanced complexity to avoid overfitting. Fine-tuning the learning rate and training threshold during model training was pivotal in achieving stable and efficient convergence. Consequently, our B-P neural network-based data asset value evaluation model, crafted with a meticulously configured multi-hidden layer structure, not only provided valuable insights for researchers but also fortified the technical foundation for practical evaluation, showcasing our expertise and commitment to excellence in neural network design and application.

4.3. Comparative Analysis of Evaluation Effects

In a meticulous examination of the efficacy of multiple linear regression, B-P neural network, and random forest models for data asset valuation, we found that machine learning models [8], particularly random forests and B-P neural networks, demonstrated remarkable superiority over traditional multiple linear regression. These advanced techniques not only surpassed the latter in terms of key metrics like root mean square error (RMSE), mean absolute error (MAE), and goodness of fit (R^2), but also displayed a heightened level of prediction accuracy and stability, highlighting their exceptional ability to grapple with the complexities of data evaluation. While the B-P neural network model also performed commendably, a nuanced observation revealed a slight deviation between its predicted results and actual values within certain specific ranges, emphasizing the need for ongoing refinement to ensure optimal performance. Overall, this analysis underscores the transformative potential of machine learning in revolutionizing the assessment of digital assets and equips practitioners with a robust set of tools to navigate the intricate landscape of the data-driven economy. In contrast, the results of the random forest model are more reliable and accurate. As for the multiple linear regression model, it has a significant deviation in predicting high-value data assets, and its performance is far inferior to the previous two.

Overall, the random forest model performs the best in evaluating the value of data assets, providing valuable references for researchers and strong support for data evaluation in practical applications. With the continuous development of machine learning technology, its application in data asset estimation will become increasingly widespread and in-depth [9].

5. Conclusion

Notably, the exclusive reliance on Youyi Data Network's data restricts dataset diversity and scope due to openness constraints, limiting the analysis's comprehensiveness. To address this and enrich our understanding, future research should incorporate diverse data sources, investigate the impact of various factors on model performance across industries and contexts, and enhance model interpretability, particularly for the random forest model, to foster wider adoption and develop more precise and effective valuation strategies for the ever-evolving landscape of data assets. As data trading platforms become more open and data-sharing practices become more common, future research should aim to integrate data from multiple sources to validate and expand the applicability of the findings comprehensively. The study primarily focuses on textual and numerical data assets, leaving the valuation of unstructured data assets like images, videos, and audio relatively unexplored. Given the proliferation of multimedia data and its pivotal role in value creation and trading, future research should intensify its focus on valuing unstructured data assets and analyzing the distinctions in their value attributes and evaluation approaches compared to other data types. Regarding modeling techniques, this paper employs machine learning models such as multiple linear regression, B-P neural networks, and random forests, with the latter emerging as the most effective for data asset valuation. However, the landscape of machine learning is constantly evolving, introducing more sophisticated methods and technologies. Future endeavors could explore the integration of diverse machine learning models or incorporate deep learning methods like convolutional neural networks (CNN) and recurrent neural networks (RNN) to enhance the precision and efficiency of data asset valuation. Indeed, exploring strategies to blend and integrate various models holds great potential in harnessing their unique advantages to deliver a more comprehensive and nuanced assessment of data asset value.

Combining the prowess of the B-P neural network's pattern recognition with the rugged stability of the random forest can enhance data asset valuation. This blend aims to transcend the boundaries of individual models, offering a comprehensive perspective that captures the intricate nuances of valuation. As researchers hone integration strategies, we anticipate the emergence of more sophisticated methods, pushing the boundaries of accuracy. The future of data asset valuation is bright, with room for growth and improvement, leveraging new technologies to deepen our understanding of data's true worth.

References

[1] Uusitalo T, Hanski J, Kortelainen H, et al. Real Value of Data in Managing Manufacturing Assets. IEEE, 2021. DOI: 10. 1007/978-3-030-64228-0_15.

[2] Jinmao L. Innovative Thinking on Development of Internal Audit of Commercial Banks in the New Era. Journal of Finance and Accounting, 2021, (3). DOI: 10. 11648/J. JFA. 20210903. 13.

[3] Shehab M F, Mohamed M. A. El-Sheikh, Hamdy M. AhmedAmina A. G. MabroukM. MirzazadehM. S. Hashemi. Solitons and other nonlinear waves for stochastic Schrdinger-Hirota model using improved modified extended tanh-function approach. Mathematical Methods in the Applied Sciences, 2023, 46 (18): 19377-19403. DOI: 10. 1002/mma. 9632.

[4] Karic K, Blagojevic M. Statistical analysis of ISO/IEC and IEEE standards in the field of artificial intelligence, machine learning and data mining. IEEE, 2021.

[5] Du Q, Zhai J. Application of artificial intelligence Sensors based on random forest algorithm in financial recognition models. Measurement: Sensors, 2024, 33. DOI: 10. 1016/j. measen. 2024. 101245.

[6] Kumar S, Garg C, Vashisht P. Method and System for Improved Consensus Using Bootstrap Resampling [P]. US202016884579. US2021374125A1 [2024-07-13].

[7] Koapaha H P, Ananto N. Bagging Based Ensemble Analysis in Handling Unbalanced Data on Classification Modeling. Klabat Accounting Review, 2021. DOI: 10. 60090/kar. v2i2. 589. 165-178.

[8] Mourad N. Robust smoothing of one-dimensional data with missing and/or outlier values. IET signal processing, 2021, (5): 15.

[9] Envelope F M O A. Multiple Linear Regression Model for Improved Project Cost Forecasting. Procedia Computer Science, 2022, 196: 808-815.