HiLL PUBLISHING

Engineering Advances, 2023, 3(2), 139-143
https://www.hillpublisher.com/journals/ea/
ISSN Online: 2768-7961

Research and System Design of Open Government Data Value Measurement System Based on Integrated Machine Learning

Wenjie Yu

Chongqing Jiaotong University, Chongqing, China.

How to cite this paper: Wenjie Yu. (2023) Research and System Design of Open Government Data Value Measurement System Based on Integrated Machine Learning. *Engineering Advances*, 3(2), 139-143.
DOI: 10.26855/ea.2023.04.012

Received: April 17, 2023
Accepted: May 15, 2023
Published: June 13, 2023

***Corresponding author:** Wenjie Yu, Chongqing Jiaotong University, Chongqing, China.

Abstract

At present, due to the data assets trading information asymmetry, in addition, the value of data assets in different application scenarios is very different, it is difficult to fully reflect the value of data assets, because the data assets is the most important in the digital economy development, the most active factors of production, and China by the new historical stage of development, in-depth analysis of data problems in the process of asset marketization, by improving the understanding of the role of data assets, activate the industry internal power to expand and strengthen the data assets, expand the effective demand and supply of data assets, develop data assets trading rules, is very necessary. This paper aims to use the data trading market in recent years brought by the rapid development of historical transaction information, in the market evaluation of the existing research, using the frontier method of artificial intelligence system to analyze the characteristics of data assets, design to evaluate the value of data assets and design the data asset pricing system.

Keywords

Asset Evaluation, Government Data, Machine Learning

1. Background

Today, the pattern of government big data, Internet big data and industry big data "dividing the world" has been in China, but as Premier Li once said: "At present, more than 80% of China's information data resources are in the hands of government departments at all levels". Therefore, the development and utilization of government big data is the key part of big data development. Then, although the government data contains huge political, economic and social value, after open to the society will also create greater public value, government data value, and measurement index has not yet, value measurement lack of scientific and applicable method, there is no rigorous specification, lead to the government data management "countless" in the statistical accounting, marketization, asset work lack of value basis.

This study for the above problems, based on the analysis of the government data value of the dimension, composition, develop the measurement method, formulate measure specification, build the government data value measurement system, the results for the government effectively and safely control and management data, further strengthen the construction of the government itself, and realize the value of government data, has far-reaching significance.

2. The economic environment of government data

Nowadays, mobile Internet, cloud computing and other technologies have penetrated into every aspect of our lives, people can obtain many aspects of data, human beings have entered a new era of creation, acquisition and use of data. According to IDC's Data Asset 2025 report, the global data volume will reach 175 ZB in 2025 (see Figure 1). For example, in the process of COVID-19 prevention and control, the government and even relevant departments have opened

the ID numbers, transportation, communications, medical treatment and other data of confirmed or suspected cases to specific groups, effectively preventing the spread of the epidemic. Promoting the opening of government data under the track of the rule of law can not only effectively accelerate the process of government data opening, ensure the good and orderly operation of the data opening system, but also effectively prevent the abuse of data, and further promote the efficient and safe circulation of data. In March 2020, the Opinions of the CPC Central Committee and The State Council on Building a More Perfect System and Mechanism for Market-based Allocation of Factors show that data, as a new factor of production, is parallel with traditional factors such as land, labor, capital and technology [1].

Open government data has many uses: open geographic data can be used to guide mining, forestry, agriculture, fishery, energy, navigation,; and open meteorological data can play a guiding role in agricultural production, disaster management and insurance forecasting, etc.

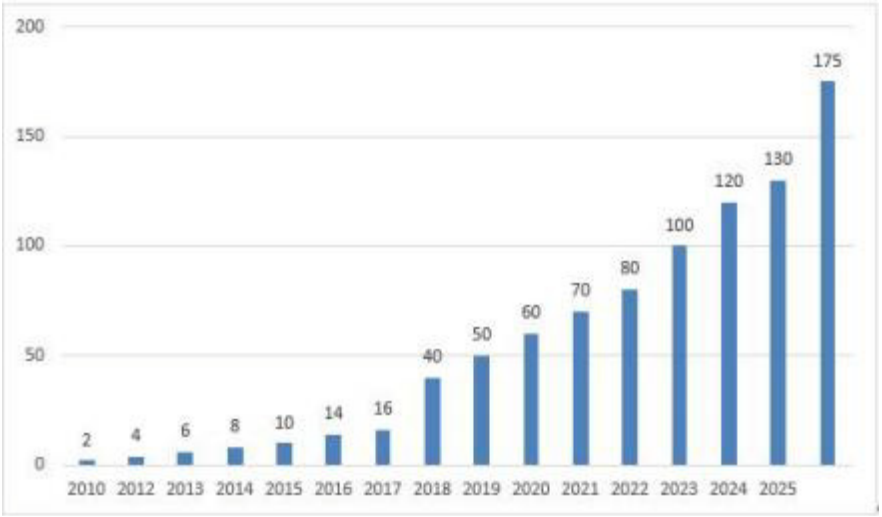

Figure 1. Amount of data assets.

3. The social environment of government data

At the government level, big data has been mentioned in the "two sessions" for four consecutive years since 2014, and in the 2017 government work report emphasized the big data strategy. On October 29,2020, approved by the "difference" planning "accelerate digital industrialization", encourage all walks of life open electricity, social, search, data, data sharing, promote the development of the third party big data service industry, and pointing to the prominent obstacles facing the current big data [2].

The White Paper on Big Data (2020) points out that with the accelerating global digital economy and the rapid development of 5G, artificial intelligence, the Internet of Things and other related technologies, data has become a key strategic resource in the new round of global competition. We should obtain and master more data resources as much as possible, and occupy a dominant position in the new round of global competition. The proposal of the 14th Five-Year Plan shows that promoting the development and utilization of data resources can promote the establishment of basic systems and norms such as property rights of data resource, cross-border transmission and security protection. Therefore, basic public data resources should be shared in an orderly manner, and a national data sharing and open platform should be built to protect national data to be safe, and further build a more perfect data governance system. The central committee of the communist party of China and the State Council on building a more perfect elements of marketization configuration system mechanism, points out to speed up the development of cultivating data market, innovative trading mechanism, improve social data resource value, pointed out the direction for the marketization of data elements configuration, to promote the number of security calculation, trusted environment data privacy protection application of new technology standards, strive to break the free flow of data barrier bottleneck [3].

In the new epidemic prevention and control, to meet the needs of using more specific cases of individual data, conducive to the public publicity, the Beijing municipal government released the patient's gender, age, living area, hospital, symptoms, trajectory, the history of contact, such as case data, contains the basic information of the cases, behavior characteristics and diagnosis and treatment, etc., this kind of more accurate government data than statistics more comprehensive, also have greater depth of use space.

4. Research core

On the basis of the OGD value measurement model by using the Starberg model, this method uses the frontier me-

thod of artificial intelligence to systematically analyze the characteristics of the value of government data assets. According to the application scenarios of the data, it is divided into different evaluation categories, which is extracted and quantified based on the characteristics of the characteristic price system, so that the characteristic system is unique and adaptable to the situation. Then, convolutional neural network under different application scenarios and other machine learning algorithms, so as to obtain the construction of open government data asset measurement model and valuation system. This study can be roughly around the "study preparation data preparation feature engineering construction model building report and recommendations".

5. Core Content

(1) Introduction of the theoretical model of the OGD value measure

1) According to the requirements of definition/index measurement, establish the measurement index system of government data value on the basis of aforementioned key influencing factors; according to the requirements of model measurement, sort out the available measurement methods, combine the research of existing data products and data assets to, the value of government data in different dimensions;

2) Study the government data value in value composition, utility, supply and demand status, data usage scenario condition change of adjustment and correction, with reference to asset appraisal, real estate appraisal specification, right, from the purpose, like, time, best use, cost, objective, value and reasonable angles, put forward the government data value measurement basic principles.

(2) Introduction of the algorithm of the OGD value measurement system

Government data is characterized by diversity, complex characteristics and high redundancy, which makes it difficult to evaluate and classify government data. In order to ensure the accuracy of data prediction, if the traditional evaluation method is adopted, the complex and tedious characteristic engineering should be carried out in the early stage, with low efficiency. And the early feature engineering will greatly affect the accuracy of the model. Therefore, we use the hybrid integrated machine learning algorithm research to innovate and improve this problem. The following machine learning integrated algorithm model is used for the early feature engineering and the valuation and pricing of data assets:

1) Convolutional neural network: CNN is a kind of feedforward neural network. Compared with other neural network learning structures, convolutional neural network is better at processing images and recognizing speech, and requires fewer parameters, so regular data can be obtained. In order to make the convolutional neural network use the two-dimensional structure of the input data, the convolutional neural network used this time is composed of multiple convolutional layers and tips, with correlation weights and pooling layers. Use the backpropagation algorithm for training.

2) XGBoost Algorithm: Use the Boosting method to improve the accuracy of the total classifier by adding multiple weak classifiers. Using the XGBoost algorithm, the modified classification regression tree is used to transform, and then a new tree is added to continuously reduce the total loss function. The traditional decision tree is split, and the principle of each step is to reduce the above loss.

3) LGB algorithm: LGB algorithm is mainly composed of three algorithms: First, similar to the weighted quantile sketch of XGBoost, the histogram algorithm is also optimized in the process of decision tree finding segmentation points to reduce the repeated traversal of a large amount of data. Specifically, the histogram algorithm first divides the continuous data of each feature into K bins, that is, a certain amount of data in each bin, from which the original continuous data becomes discrete bin data

Second, after using the Histogram algorithm, it was further optimized using LightGBM. Because of the large amount of data, it is not suitable to use decision tree growth strategies, so most GBDT tools are abandoned. However, the leaf growth algorithm is used, which is more suitable for this data processing because of the depth limitation. For multi-threading optimization, controlling model complexity and preventing overfitting, using XGBoost's Level-wise growth strategy, which traverses the data can split the leaves of the same layer simultaneously.

The l-wise growth strategy, which can simultaneously split the leaves of the same layer. Third, the one-sided gradient sampling algorithm called Gradient-based One-Side Sampling. Using this algorithm, we can effectively reduce samples, exclude most samples with small gradients, and calculate the information gain with only the remaining samples, which can reduce the amount of data and guarantee the accuracy.

4) Random forest: Random forest algorithm has the basic characteristics of Bagging algorithm, which is its expansion. On the basic learning of decision making, Bagging integration is constructed to further introduce random attribute selection in the process of training the decision tree. That is, when choosing the partition attributes, the traditional decision tree obtains one of the most attributes in the base node. However, using the random forest algorithm, for each node of the base decision tree, a subset of k properties is randomly selected from the attribute combination of the node, and then an optimal attribute user partition is selected from this subset. The random forest is used to test whether the division of subsets is too large. The random forest model is a model based on variance, so the predicted result is a good

model used to test the selection of feature subsets [4].

(3) System design

The value of research and design of government data should not only lie in the establishment of government data asset evaluation model, but also promote the development of government data and even the government itself. Reflected in the following aspects:

1) More clearly divides the use scenarios of government data. Because government data has multiple characteristics, it can be used in different scenarios. For the same government data, different values show different values. Multiple application scenarios of government data assets have been determined through the methods of literature learning and expert interview. Before pricing government data assets, Python is first used to conduct text similarity analysis of the titles of open government data, and divide them into different categories according to the different usage scenarios.

2) Promote the rationalization and development of government data asset pricing. According to the existing open government data, construct appropriate index system and form the evaluation model framework; the government data asset valuation pricing model based on mathematical analysis methods such as Starkberg model, provide more and more widely accepted practical basis to promote the fair and reasonable realization of government data asset pricing.

3) In order to support the real-time data valuation calculation, the real-time valuation system of government data assets is established by using machine learning related algorithms based on the research of the government data asset measurement model. Improve the valuation efficiency by tuning the model parameters rate. The data valuation system mainly includes: data collector, data processor, interactive module and data valuation model. The system is composed of five modules: data input interface module, data acquisition module, data preprocessing module, feature engineering construction module, and valuation and pricing module.

S1 Data input interface module

In view of the multiple characteristics and multi-dimensional characteristics of open government data information, the first module designed for the evaluation system is the data input interface module. This module is composed of two submodules: input and classification. The data input module is the first sub-module, which is responsible for the input of 7 independent variable data, including data size, data type, number of fields, number of data bars and acquisition time. The classification module is another sub-module. According to more than 10 usage scenarios such as scientific research technology, industrial economy, economic development, financial credit investigation, public opinion monitoring, etc. After text similarity analysis, the input data entered in the first submodule is divided into different comparison schemes.

S2 Data acquisition module

Government data assets have the characteristics of large data quantity and diverse data types. Therefore, the data acquisition module should be established specifically. First of all, the Python is used to crawl the information of the open government data platform to the maximum extent, and to crawl the asset title, data type, collection time, field number, data size, the number of data bars and other information types of the block data and API government data transaction.

S3 Data preprocessing module

In view of the problems of government data, such as complicated data types, numerous data sources, and possibly missing data, the system sets up an abnormal data pre-processing module to improve the data quality. After collecting and storing the data, the module is called to input the data asset history information obtained in the module to conduct consistency test, regression analysis, missing data processing and other pre-processing operations.

S4 Feature engineering building module

It will be a number of relevant information input feature engineering construction modules of open government data, which can output the feature data corresponding to the input. This module is specifically implemented:

(1) Text analysis: First, the text title of the collected relevant government data assets is transformed into word vector, then the word vector is averaged, and the average model is further transformed into sentence vector. Finally, the characteristic dimension reduction of the title of the open government data assets is realized through convolutional neural network (CNN).

(2) One-hot code: One-hot code should be used to extend the value of discrete features to Euclidean space, so that the distance between data type extraction and classification features is more reasonable [5];

S5 Valuation and pricing module:

The function of this module is to input characteristic information and obtain the reference pricing of the estimated government data assets, achieving the following purposes:

(1) OGD value measurement theoretical model: here we should input m comparable examples, and then encode the characteristic data of government data assets to obtain the characteristic price matrix of $m \times n$

(2) OGD value measurement algorithm system: establish government data pricing model based on machine learning algorithm. According to the results of OGD value measurement theory model and other information into the feature data of comparable examples, various machine learning algorithms such as XGboost, LGB and CB are used to train the government data, and the integrated model with strong regression ability and good robustness is synthesized. The resulting comparable cases are input into the model to accurately calculate the value of the government data assets to be

estimated.

The system can realize: data price information perspective. Through the user entering the data title, data type, data set size, collection time, data bar number and selecting data tags, the integrated machine learning model is used to evaluate the data asset price and conduct batch data pricing, and effectively predict the data by importing csv, text, and xlsx format files. Reasonable balance of some problems caused by information asymmetry, forming the principle of fair trade.

References

[1] Deng Shaocun, Zhang Jianqin, Zhang Xuedong, et al. Research and development of visualization system for COVID-19 based on WebGIS [J]. Journal of Beijing Jianzhu University, 2020, 36 (4): 7.

[2] Cao Hongming, Xu Xiaolan, Jia Wenqin, et al. Research and Reflection on Promoting the Development of the Sharing Economy and Cultivating New Driving Forces for Economic Development [J]. China Development, 2018, 18 (1): 5.

[3] None Opinions of the Central Committee of the Communist Party of China and the State Council on Building a More Perfect System and Mechanism for Factor Marketization Allocation: Promoting the Autonomous and Orderly Flow of Factors and Improving the Efficiency of Factor Allocation [J]. Chinese Talent, 2020.

[4] Li Tingyang, Luan Xin, Peng Zhenghong. Application of Decision Tree Learning Algorithm in Traffic Mode Selection Model [J]. Journal of Wuhan University: Engineering Edition, 2013, 46 (3): 5.

[5] Tang Peng. LDA based feature extraction and its application in facial recognition [D]. Hebei University.

[6] Feizollahi, M. J., et al. (2018). A novel approach to measure open data value based on machine learning. Government Information Quarterly, 35(3), 354-363. doi: 10.1016/j.giq.2018.05.003.

[7] Kodama, F., et al. (2019). An integrated approach for measuring open data utilization. Online Information Review, 43(7), 1169-1184. doi: 10.1108/OIR-01-2019-0020.

[8] Zhao, J., et al. (2020). Measuring the public value of open government data: A machine learning-based approach. Journal of the Association for Information Science and Technology, 71(4), 447-459. doi: 10.1002/asi.24245.

[9] Moon, M. J., et al. (2021). Building a machine learning-based open government data value measurement framework. Government Information Quarterly, 38(1), 101470. doi: 10.1016/j.giq.2020.101470.

[10] Bai, X., et al. (2022). Integrating machine learning techniques into open government data value measurement. Journal of Systems Science and Systems Engineering, 31, 23-42. doi: 10.1007/s11518-021-5504-6.