DOI: 10.54254/3029-0880/1/2023006

Research on the data platform needs of university faculty and students

Xinping Lv

Shandong Normal University, 250399

15854585772@163.com

Abstract. In today's world, big data permeates every industry, leading to a substantial increase in research data and driving the scientific paradigm towards data-intensive research. With the development of intensive research, the demand for data processing and analysis is growing, particularly pronounced in the fields of research and management in universities. Currently, education in China is transitioning from informatization to digitization to enhance the quality of education, research by faculty and students, and the overall operational quality of schools, thereby achieving high-quality development in higher education. Digitization has become a crucial innovative path for the modernization of education in China and is an important strategy for building a high-quality education system. However, many university teachers and students still face challenges such as time-consuming and labor-intensive data processing, lack of relevant technical expertise, and difficulty in seeking guidance from professional experts. At the same time, the number of professionals in the big data industry is gradually increasing, and they aspire to enhance their professional skills through practical experience, earn additional income, but lack suitable avenues. To address these market challenges and pain points, fill market gaps, promote the academic and research capabilities of university faculty and students, improve personnel management efficiency, and advance scientific research in China, this survey is conducted to propose practical and feasible suggestions and actively engage in implementation.

Keywords: data, service platform, university faculty and students, innovative approaches, solutions

1. Research overview

With the rapid development of the era, the importance of data to humanity has gradually increased. In the present age, the value of data has been preliminarily tapped, yielding initial results in areas such as

DOI: 10.54254/3029-0880/1/2023006

national defense, livelihood, economy, and politics. [1] As the most active technological innovation factor in the new round of industrial revolution, big data is comprehensively reconstructing global fields such as production, circulation, distribution, and consumption, exerting a comprehensive and profound impact on global competition, national governance, economic development, industrial transformation, and social life. In China's economic digitization and transformation, big data is expected to play an even more important innovative role. Big data will continue to create higher value and integrate more extensively into various aspects of people's lives in the future.

However, many university teachers and students currently face challenges such as time-consuming and labor-intensive data processing, a lack of relevant technical expertise, and difficulties in seeking guidance from professional experts. At the same time, the number of professionals in the big data industry is gradually increasing, and they aspire to enhance their professional skills through practical experience, gaining additional income, but lack suitable avenues.

In summary, to address these market challenges and pain points, fill market gaps, promote the academic and research capabilities of university faculty and students, improve personnel management efficiency, and advance scientific research in China, this paper proposes a study on the influencing factors of the intention to choose data processing solutions among university faculty and students nationwide. Following the analytical logic of "behavioral intention—usage behavior—usage performance" in the selection of data processing solutions by university faculty and students, this research combines relevant literature to study theories and models. By incorporating the current development status of data processing solutions for university faculty and students and interview results, a thorough investigation into the selection of data processing solutions is conducted.

By presenting hypotheses and establishing theoretical models, and devising a survey plan based on the actual situation, data is collected through questionnaires according to model variables. After analyzing and processing the data, empirical verification of the research hypotheses is conducted. Based on the analysis of the data, conclusions are drawn, and, finally, substantive recommendations are made for the selection of data processing solutions by university faculty and students based on the conclusions of the above issues.

2. Model construction and analysis

2.1. Building a predictive model for user processing needs based on sampling stepwise regression

2.1.1 Model Construction. Initially, we explore the relationship between the data processing needs of different users and their individual attributes. The data processing needs of different users are influenced by their individual attributes. When constructing the model, a selection of universities nationwide is taken as the primary indicator, and the individual attributes of faculty and students at various universities are statistically recorded.

Let the individual attributes of faculty and students at various universities be denoted as the matrix X=(xij). The data processing demand rate of faculty and students at various universities is denoted as y_i, where i represents the user's identity, and j represents the user's major. The function relationship

DOI: 10.54254/3029-0880/1/2023006

between the data processing needs of different users (dependent variable) and their individual attributes (independent variable) is constructed as shown in Equation (1).

$$y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{122} x_{i,122} + \varepsilon_i \tag{1}$$

Where, α represents the intercept, ε is the residual term, and βj represents the regression coefficient of the jth variable in the regression equation.

2.1.2. Model analysis. Next, we solve Equation (1) based on the Spike-slab sparse function and stepwise regression to build a variable selection model. Due to the presence of multicollinearity among variables and the large number of independent variables in the equation, ordinary least squares (OLS) cannot precisely estimate the coefficients of each independent variable. [2] Simultaneously, we aim to retain all influencing factors in the model unless a variable has no impact on the user's data processing needs. The specific steps are as follows:

(1) Randomly sample the independent variable X based on the Spike-slab sparse function, obtaining a sample denoted as Xr.

(2) Build a stepwise regression model, regressing y on Xr. Obtain regression results at a specified significance level (significance level set to 0.05).

(3) Repeat steps 1 and 2 a sufficient number of times, take the average of all validated regression results, and obtain estimates for all parameters in the equation.

Here are the specific steps. Firstly, let the coefficients of the independent variable X be a column vector $\beta = (\beta_j)$. Based on β, construct $\gamma = (\gamma_j)$, where, when $\beta_j = 0$, let $\gamma_j = 0$; conversely, when $\beta_j \neq 0$, let $\gamma_j = 1$. Usually, γ can be constructed based on the Bernoulli distribution, as shown in Equation (2):

$$\gamma \sim p_j^{\gamma_j} \left(1 - p_j\right)^{1-\gamma_j} \tag{2}$$

Where, for each stepwise regression model, when we want the expected number of variables to be m, we can set $p_j = \frac{m}{n}$, where n represents the total number of independent variables. Then, based on the formula, sample a prior γ, and then, according to $\gamma_j = 1$ select the corresponding variable x_j with $\beta_j \neq 0$, denoted as the set X_γ, which is the current stepwise regression sample of independent variables.

Next, build the stepwise regression model and estimate the values of the parameters in the model $y = f(X_\gamma)$. Obtain the fitted values of the regression coefficients β and α corresponding to the current X_γ. The stepwise regression method ensures that the variables retained in the model are those with a significant impact on the dependent variable and eliminates multicollinearity.

DOI: 10.54254/3029-0880/1/2023006

Finally, repeat sampling a sufficient number of times to ensure model convergence. Let the regression parameters of the ith stepwise regression be denoted as $\varphi^{(i)} = (\alpha, \beta)^{(i)}$, and a series of fitting results $(\varphi^{(i)})$. Calculate the mean of all regression coefficients, use it as the final estimated coefficients for the independent variables, denoted as $\bar{\varphi} = \sum_{i=1}^{N} \varphi^{(i)} / N$. Thus, the relationship model between the inflow and outflow of public bicycles at point i and the land attributes of that point is expressed in Equation (3).

$$\hat{y}_i = \bar{\alpha} + \bar{\beta}_1 x_{i,1} + \bar{\beta}_2 x_{i,2} + \cdots + \bar{\beta}_{122} x_{i,122} \tag{3}$$

2.1.3. Results Analysis. Finally, after calculation, the regression coefficients of the independent variables (individual attributes) on the dependent variable (data processing demand) are obtained, as shown in the table below:

Table 1: Regression coefficients

	Collection	Preprocessing	Analysis	Visualization
Gender	0.0046	0.0052	0.0048	0.0051
Identity	0.2026	0.2068	0.2113	0.2097
Major	0.1926	0.1879	0.1903	0.1968
Degree	0.2268	0.2158	0.2214	0.2256

From the above table, the regression equations for the dependent variable and the independent variables can be obtained as follows:

y1=0.0046x1+0.2026x2+0.1926x3+0.2268x4

y2=0.0052x1+0.2068x2+0.1879x3+0.2158x4

y3=0.0048x1+0.2113x2+0.1903x3+0.2214x4

y4=0.0051x1+0.2097x2+0.1968x3+0.2256x4

From the regression equations, it can be observed that user identity, user's major, and the user's highest degree pursued are positively correlated with the data processing demand rate. Among them, the impact of the user's highest degree pursued on the data processing demand rate is the most significant, followed by user identity, and finally the user's major. Users with higher degrees have a higher demand for data processing, and teachers often have a higher data processing demand rate than students.

2.2. Building a predictive model for user platform usage intentions based on K-means clustering and negative binomial model

According to the survey results, the willingness to use data service platforms by users is influenced by three factors: individual attributes, data processing needs, and challenges in data processing needs. Therefore, this study considers these three indicators as primary indicators affecting the willingness to use data service platforms and investigates their impact on people's willingness to use data service

platforms. [3] Specifically, four secondary indicators representing individual attributes, five representing data processing needs, and six representing challenges in data processing needs are selected, as shown in the table below:

Table 2: Primary and secondary indicators

Primary Indicators	Secondary Indicators
Individual Attributes	Gender, Identity, Major, Degree
Data Processing Needs	Data Collection, Data Preprocessing, Data Analysis, Data Visualization, Data Mining
Data Processing Challenges	Time-consuming, Lack of Professional Knowledge, High Abnormality in Data, Tedious Tasks, Inaccuracy in Collected Data, Existence of Factual Bias in Data

2.2.1. K-Means Clustering. K-means clustering is a commonly used sample-based clustering algorithm. Its basic idea is to pre-specify the number of categories, k, and iteratively update cluster centers and partitions by minimizing a loss function to select the optimal partition. In the application of this method, for a dataset X of n sample data for university faculty and students, each sample is composed of a feature vector with m attributes, i.e., X={x1, x2, …, xn}.

When the willingness level for data service platforms is set to k, the n samples can be divided into k subsets, C={C1, C2, …, CK}. The calculation steps using the k-means clustering method are as follows:

(1) Initialization: Randomly select k sample data h (0) from the faculty and student dataset X as the initial cluster centers for the willingness level k. h (0)={h1 (0), h2 (0), …, hk (0) | hi (0)∈X, i=1, 2, …, k}, where each sample in h (0) uniquely corresponds to the initial cluster center for a user's willingness level.

(2) Cluster the samples: Calculate the distance from (n-k) samples to the cluster centers:

$$\mathrm{d}(x_j - \mathrm{h}_l^{(0)}) = \sum_{w=1}^{m} (x_{wj} - h_{wl}^{(0)})^2 \tag{4}$$

Where, Xwj is the wth attribute value of sample xj, and hwl (0) is the wth attribute value of sample hl (0). xj∈X, hl (0)∈h (0). Assign each sample to the willingness level category of its nearest cluster center, resulting in the clustering result: C={C1 (0), C2 (0), …, CK (0)}.

(3) Calculate new cluster centers: For the initial clustering result C (0), calculate the mean of the samples contained in each willingness level, and use it as the new cluster centers h (1) = {h1 (1), h2 (1), …, hk (1)}.

$$h_i^{(1)} = \frac{1}{|C_i^{(0)}|} \sum_{x \in C_i^{(0)}} x \tag{5}$$

Where, |Ci(0)| is the number of samples in the initial clustering result for willingness level i, i=1, 2, …, k.

DOI: 10.54254/3029-0880/1/2023006

(4) Iterative optimization: The k-means algorithm uses the sum of squared errors criterion function to evaluate clustering performance. The formula for calculating the squared error E of the final clustering result is as follows:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \| x - \mu_i \|^2 \tag{6}$$

Where, $\mu_i = \frac{1}{|C_i|} \sum_{c \in C_i} x$ is the mean vector of the final clustering result for category Ci, and E reflects to some extent the closeness of the samples around their mean vectors for different willingness level categories. A smaller E indicates higher similarity among samples in different willingness level categories.

To minimize the squared error, iteration is usually adopted: repeat the above steps, and after t rounds of iteration, if the willingness level corresponding to each sample in dataset X is the same as the result of the (t-1) round of iteration, stop the iteration and let:

$$C = C^{(t)} \tag{7}$$

Where, C(t) is the clustering result obtained after t iterations.

2.2.2. Model Establishment. Next, we construct a multifactor influence model to investigate the correlation between the willingness to use data service platforms and various influencing factors. Since the dependent variable (i.e., willingness to use data service platforms) is discrete count data, we consider using a negative binomial regression model to build this influence factor model. [4] Firstly, we use the negative binomial regression model to establish the functional relationship between the willingness to use data service platforms and different influencing factors, as shown in the equation:

$$\ln(y_{i,t}) = \alpha + \beta_1 x_{i,t,1} + \beta_2 x_{i,t,2} + \beta_3 x_{i,t,3} + \cdots + \beta_{38} x_{i,t,38} + \varepsilon_{i,t} \tag{8}$$

Where yi, t represents the willingness to use data service platforms for user i with attribute t, i represents the user, t represents the user attribute, xi, t, j represents the jth factor influencing shared bike demand, βj represents the regression coefficient of the jth independent variable in the regression equation, α represents the intercept, and εi, t represents the residual term.

2.2.3. Results Analysis. Based on questionnaire data and computation results, we defined the willingness of users to use data service platforms k into 6 levels, represented by numbers 1-6. Level 1 indicates the strongest willingness, decreasing in sequence. Levels 1-2 are defined as strong willingness, levels 3-4 as moderate willingness, and levels 5-6 as low willingness.

From clustering operations and regression equations, we can determine the strength of the impact of three factors, namely, user individual attributes, data processing needs, and challenges in data processing needs, on user willingness to use data service platforms. The table below illustrates the strength of these impacts:

DOI: 10.54254/3029-0880/1/2023006

Table 3: Impact strength table

Individual Attribute	Impact Level	Data Requirement	Impact Level	Demand Challenges	Impact Level
Gender	6	Data Collection	2	Time-consuming and Labor-intensive	2
Identity	4	Data Preprocessing	4	Lack of Expertise	4
Major	4	Data Analysis	1	High Abnormality	5
Degree	2	Data Visualization	1	Tedious Tasks	3

From the charts, it can be observed that among individual attributes, the user's degree is the main influencing factor for the willingness level. Users with a demand for data analysis and data visualization show a higher willingness to use data service platforms. When data processing tasks are excessively time-consuming and labor-intensive, users are willing to use platform websites to address existing issues.

3. Conclusion and Recommendations

3.1. Conclusion

The lack of standardized data and the absence of unified data standards have led to poor data quality and a lack of effective means for obtaining high-quality data. With the improvement of overall school capabilities, there is a higher demand for data processing. The experiential quality of data service platforms and the stability of the platform are the main factors influencing the participation of teachers and students in data service platforms. [5] Integrating ordinary big data service platforms with task posting and bidding is a good way to enhance the experience of data service platforms. The market potential for data service platforms is enormous.

3.2. Recommendations

(1) Establish a new type of data service platform to address the data processing needs of university faculty and students.

(2) Provide personalized services to offer a precise and comprehensive platform experience for the customer base.

(3) Innovate the platform service model, creating an innovative development model centered around task posting and bidding.

(4) Implement a "mutual benefit" model.

(5) Emphasize the regulatory effectiveness of the platform, ensuring the security of the platform's information system.

(6) Strengthen data management.

References

[1] Wu, G., & Chen, G. X. (2018). Operational Mechanism of Big Data Governance in Universities:

Functions, Issues, and Improvement Strategies. University Education Science, 0(6), 34-38.

[2] Yang, Y. (2020). Planning, Design, and Implementation of Big Data Platforms in Universities. Journal of Shenzhen University: Science and Engineering, 37(S01), 146-149.

[3] Hu, S. X., Jing, Z., & Wang, H. J. (2022). Research on Key Elements and Optimization Paths of Big Data Governance System in Chinese Universities: A Research Perspective Based on DEMATEL-ISM. Educational Research on Electronics, 43(11), 38-44+52. DOI:10.13811/j.cnki.eer.2022.11.005.

[4] Using DAF and AIDA to Scope Data Management Needs [EB/OL]. (2018-11-21). http://www.data-audit.eu/docs/DC101_DAFAIDA_150709.pdf

[5] Jones, S., Ball, A., & Ekmekcioglu, C. (2008). The Data Audit Framework: A First Step in the Data Management Challenge. International Journal of Digital Curation, 3(2), 112-120.