future internet

MDPI

Article

Financial Data Quality Evaluation Method Based on Multiple Linear Regression

Meng Li, Jiqiang Liu * and Yeping Yang

School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China; 18112051@bjtu.edu.cn (M.L.)
* Correspondence: jqliu@bjtu.edu.cn

Abstract: With the rapid growth of customer data in financial institutions, such as trusts, issues of data quality have become increasingly prominent. The main challenge lies in constructing an effective evaluation method that ensures accurate and efficient assessment of customer data quality when dealing with massive customer data. In this paper, we construct a data quality evaluation index system based on the analytic hierarchy process through a comprehensive investigation of existing research on data quality. Then, redundant features are filtered based on the Shapley value, and the multiple linear regression model is employed to adjust the weight of different indices. Finally, a case study of the customer and institution information of a trust institution is conducted. The results demonstrate that the utilization of completeness, accuracy, timeliness, consistency, uniqueness, and compliance to establish a quality evaluation index system proves instrumental in conducting extensive and in-depth research on data quality measurement dimensions. Additionally, the data quality evaluation approach based on multiple linear regression facilitates the batch scoring of data, and the incorporation of the Shapley value facilitates the elimination of invalid features. This enables the intelligent evaluation of large-scale data quality for financial data.

Keywords: quality evaluation; analytic hierarchy process; multiple linear regression; index system; Shapley value

Citation: Li, M.; Liu, J.; Yang, Y. Financial Data Quality Evaluation Method Based on Multiple Linear Regression. *Future Internet* **2023**, *15*, 338. https://doi.org/10.3390/fi15100338

Academic Editor: Massimo Cafaro

Received: 20 August 2023
Revised: 24 September 2023
Accepted: 28 September 2023
Published: 14 October 2023

1. Introduction

Quality evaluation commonly pertains to the methodologies that organizations employ to ensure the attainment of strategic objectives, establish regulatory compliance, and monitor data integrity [1]. In the past, data collected from various transaction systems within companies or institutions were often regarded as a mere by-product of business operations, possessing limited value beyond the transactions [2]. Today, as the complexity and volume of data continue to escalate, numerous companies and businesses are compelled to adapt their business models accordingly. Customers need personalized products, and service products must be industrialized. These factors inevitably influence business processes and organizational strategies, and high-quality data are a prerequisite for meeting these changing business demands and achieving corporate agility objectives [3]. A single data management solution is insufficient to meet the demands of the business. Therefore, different approaches to integrating solutions to data problems must be employed [4].

Financial institutions must ensure the completeness and accuracy of their data quality so that regulatory bodies can better understand the operation of the institutions, predict risks, and take corresponding measures [5]. In the era of rapid big data technology advancement, financial institutions have been compelled to enhance their management, processing, and analysis of vast volumes of personal customer information. This imperative arises from the goals to enhance business efficiency and strengthen risk management capabilities. Institutions are increasingly utilizing machine learning technologies to analyze customer behavior, predict market trends, and identify fraudulent behaviors. These technologies also must be able to support a large amount of high-quality data; therefore, financial institutions must establish effective data governance and evaluation mechanisms to ensure data quality.

Mcgilvray [6] introduced a methodical yet adaptable framework consisting of ten steps to address an organization's business needs and data quality issues. Within this framework, data quality dimensions are employed to precisely define, measure, and effectively manage the quality of data. Omara et al. [7] presented a comprehensive compilation of prevalent data quality dimensions, encompassing accuracy, completeness, consistency, and timeliness. Furthermore, they proposed a novel approach to gauge row completeness utilizing a data mining model developed on neural networks. Peltier et al. [8] measured data quality from a systems perspective with more dimensions and considered customer data quality to facilitate the development of personalized interactive marketing initiatives. Taleb et al. [9] developed a scheme to evaluate data quality on large datasets using sampling strategies, which adopted some data quality measurement metrics tailored to specific data quality dimensions. Juddoo [10] conducted an investigation on a variety of data quality metrics, with a specific focus on the measurement and quantification of particular dimensions like completeness and consistency. Recently, the analytic hierarchy process (AHP) was used to integrate various data quality dimensions as well as expert preferences for quantifying a comprehensive score of data quality [11], and has been applied to healthcare [12], financial statements [13] and cloud computing [14].

For the data quality evaluation system, most experts only consider qualitative or quantitative situations, citing a single data quality evaluation method, such as fault tree analysis, AHP, gray relational degree analysis, etc. Some scholars combine commonly used engineering evaluation methods to build evaluation models on the basis of combining qualitative and quantitative methods. Through research and combining, this paper finds that many scholars have not unified the selection of data quality indicators, and most of them use more traditional methods, which lack certain innovation. This leads to the main contributions of this paper as listed below:

1. We proposed a new system for constructing quality evaluation indicators by surveying the current status of data quality evaluation research. This system forms primary indicators according to completeness, accuracy, timeliness, consistency, uniqueness, and compliance. We used AHP to determine the weight among these indicators.
2. We proposed a data quality evaluation method based on multiple linear regression (MLP) to objectively evaluate data quality. In addition, we employed et al.Shapley value for feature selection to reduce the complexity of the model and shorten the training time.
3. By using customer-related table information from a specific financial institution as an example, we verified the feasibility and accuracy of the quality evaluation system and the quality evaluation model. Additionally, we found that the evaluation and processing of incremental customer form information could be automated.

The remainder of this paper is organization as follows: In Section 2, we formulate the primary indicators and adopt the AHP to determine the weights of the indicators. In Section 3, we construct a quality evaluation model using MLP algorithms. We next utilize a concept of the Shapley value to evaluate the contribution of each indicator to the scoring model, facilitating feature selection. In Section 4, we use the customer information system of a trust company as an example for empirical analysis. In Section 5, we summarize the results of this study and discuss its contributions.

2. Construction of Quality Evaluation Indicator System

The construction of the evaluation system should consider the following points data [15]:

1. Quality evaluation focuses on each dataset as the target. Every department is responsible for managing multiple datasets [16]. These datasets are typically represented as information source tables.
2. Data usability is the fundamental prerequisite for effective data utilization. The data resource centralization department can contribute by identifying factors that impede

the basic usability of the data. These factors include empty data items, irregular formats, incomplete data, and other similar issues (Table 1).

3. Data flexibility takes center stage as a crucial consideration. When evaluating the quality of customer data types, it is essential to consider its unique characteristics while maintaining a certain level of flexibility. For instance, empty fields may be the result of non-required fields [17].

Table 1. Problem situations and examples of some data.

Problem Situation	Description and Examples
field empty	Empty values appear in required fields such as gender, nationality, address, contact information, and certificate validity period
irregular format	The 18-digit format that does not meet the standard, such as ID number, etc.
consistency error	Formats that do not meet the standard, such as the ID number does not match the gender or the date of birth, etc.
timeliness error	The validity period of the certificate has "valid until less than the current date", etc.
data record duplication	The same data are repeatedly recorded, and there is a situation of "multiple compilations for one household"

Because of multiple systems and vast amounts of data, it is crucial to strike a balance between quality and efficiency of the evaluation process [18]. To address this, a quality evaluation index framework for trustworthy data should be established, consisting of six dimensions: completeness, accuracy, timeliness, consistency, uniqueness, and compliance.

2.1. Basic Attributes of Quality Evaluation

A completeness evaluation encompasses three secondary indicators: record completeness, attribute completeness, and value completeness. The following fields should be assigned:

1. Required fields: These fields must be filled in accordance with the business rules of each department or according to the specifications outlined in the data dictionary.
2. Key fields: These fields serve as unique primary keys or facilitate associations with related data tables. Examples of such key fields include the identity card number, the unified social credit identifier, and other identification numbers or codes that act as primary keys in related associations.

Data accuracy encompasses three aspects: logical accuracy, value accuracy, and conceptual accuracy (Table 2). If the identity card number for a piece of data is not accurate or does not correspond to the name, this identity card number is considered to be one piece of erroneous data.

Table 2. Some types of data accuracy and related descriptions.

Question Type	Problem Statement
False	The data are not true, such as the ID number being non-existent
inconsistent information	The related information does not correspond, such as the ID number not corresponding to the name, or there being a situation where one ID number corresponds to multiple names
timing relationship error	The relative timing between data is not correct, such as the creation time of individual customers being later than the modification time
invalid data	Data records should not be valid relative to a certain period of time or a certain point in time; for example, the establishment time of the enterprise is later than the current time

The timeliness requirement differs depending on the update frequencies of various information resource tables. As a result, rules can be established differently [19].

A consistency evaluation measures the level of data association, assessing the logical accuracy and completeness of the same information subject across different datasets [20].

A uniqueness evaluation gauges the extent of data duplication. For a dataset that consists of four fields, customer number, customer name, customer status, and gender, if N data records have the same values for all four fields, it would produce N-1 duplicate data records. As presented in Table 3, we identified a total of two duplicate data records.

Table 3. Example of duplicate data records.

Data Record	NO.	Customer Number	Customer Name	Gender
1	1000103445	112803	He Zhixiong	2
2	1000103445	112803	He Zhixiong	2
3	1000103445	112803	He Zhixiong	2
4	1000083575	112803	Zhu Hongyan	1

Compliance ensures that the data types and precision of each data field fall within the specified range. For those fields governed by industry unified standards, adherence to the relevant standards is required. In the absence of unified standards, compliance is based on the data dictionary or relevant regulations provided by each department.

2.2. Determination of Indicator Weights Based on the AHP

The AHP is characterized by its ability to convert judgments into comparisons of importance between pairs of several factors [21]. The complete process involves decomposition, judgment, synthesis, and other steps, effectively addressing the limitations of alternative methods that aim to minimize subjective judgment from decision makers.

2.2.1. Single Scoring

To score a dataset from an information resource table, the scores of each secondary indicator can be computed using the previous scoring rules. The total score of the evaluated dataset can then be calculated by taking the weighted average.

$$Y = \sum_{n} \lambda_n X_n \tag{1}$$

where Y is the total score of the evaluated data set, λ_n is the indicator weight, X_n is the indicator score, and n is the secondary indicator number.

In the evaluation of a data set from an information resource table, the scoring range for each indicator ranges from 0 to 100 points. The scoring rules and formula are detailed in Table 4.

Table 4. Scoring rules and scoring formulas for data quality evaluation indicators.

Indicator Number	Indicator Name	Scoring Formula and Instructions
$B1$	integrity	$X_{B1} = \lambda_{C11}X_{C11} + \lambda_{C12}X_{C12} + \lambda_{C13}X_{C13}$
$C11$	record integrity	$X_{C11} = 100 \times (1 - N_{C11}/W_{C11})$ N_{C11}: the number of missing data items in the data table W_{C11} : the total number of data items that should be recorded in the data table
$C12$	attribute integrity	$X_{C12} = 100 \times (1 - N_{C12}/W_{C12})$ N_{C12}: the number of missing attributes in the data table W_{C12}: the total number of attributes that should be recorded in the data table

Table 4. *Cont.*

Indicator Number	Indicator Name	Scoring Formula and Instructions
*C*13	value integrity	$X_{C13} = 100 \times (1 - N_{C13}/W_{C13})$ N_{C13}: missing required fields in the data table W_{C13}: the total number of data items that should be required in the data table
*B*2	accuracy	$X_{B2} = \lambda_{C21}X_{C21} + \lambda_{C22}X_{C22} + \lambda_{C23}X_{C23}$
*C*21	record accuracy	$X_{C21} = 100 \times (1 - N_{C21}/W_{C21})$ N_{C21} : the number of inaccurate data items recorded in the data table W_{C21} : the total number of assigned data items in the data table
*C*22	value accuracy	$X_{C22} = 100 \times (1 - N_{C22}/W_{C22})$ N_{C22}: number of data items in the data table with inaccurate values W_{C22}: the total number of data items that should be recorded in the data table
*C*23	conceptual accuracy	$X_{C23} = 100 \times (1 - N_{C23}/W_{C23})$ N_{C23} : the number of conceptually inaccurate data items in the data table W_{C23}: the total number of data items that should be recorded in the data table
*B*3	timeliness	$X_{B3} = X_{C31}$
*C*31	data update timeliness	$X_{C31} = 100 \times (1 - N_{C31}/W_{C31})$ N_{C31}: the number of data items that are not updated in time in the data table W_{C31}: the total number of data items that should be recorded in the data table
*B*4	consistency	$X_{B4} = X_{C41}$
*C*41	logical consistency	$X_{C41} = 100 \times (1 - N_{C41}/W_{C41})$ N_{C41} : the number of data items in the data table that do not satisfy the logic W_{C41} : the total number of data items that should be recorded in the data table
*B*5	uniqueness	$X_{B5} = X_{C51}$
*C*51	record uniqueness	$X_{C51} = 100 \times (1 - N_{C51}/W_{C51})$ N_{C51}: the number of data items repeated in the data table W_{C51}: the total number of assigned data items in the data table
*B*6	compliance	$X_{B6} = X_{C61}$
*C*61	data format compliance	$X_{C61} = 100 \times (1 - N_{C61}/W_{C61})$ N_{C61}: number of data items with non-compliant format in the data table W_{C61}: the total number of assigned data items in the data table

2.2.2. Comprehensive Scoring

Trusts continuously update the collected data. Therefore, if it is necessary to showcase the overall quality of a data set from a specific information resource table over a certain time frame, the numerous evaluation outcomes within this period can be comprehensively processed:

$$S_i = \sum_k \frac{Y_k}{K_i} \tag{2}$$

where Y_k represents the single data quality evaluation score of the information resource data table, k is the order of evaluation, K_i represents the total number of times the data table is evaluated during the evaluation period, and i is the number of each information resource table.

3. Construction of Quality Evaluation Model Based on Multiple Liner Regression

In this study, we present a machine learning-based data quality evaluation method. The method revolves around the creation of a data quality evaluation indicator system. The key steps of this approach are as follows:

1. Select relevant customer information data from the customer information database, and preprocess the data based on the proposed label generation algorithm.
2. Extract quality evaluation indicators and associated data, constructing a quality evaluation indicator system suitable for machine learning models.
3. Train and test the machine learning model to complete its construction.
4. Evaluate the results using performance indicators.

3.1. Data Acquisition and Indicator Extraction

As shown in Figure 1, we extracted the table being evaluated from the ECIF (Enterprise Customer Information Facility) database by employing the data quality evaluation indicators. Subsequently, to facilitate the implementation of the machine learning model, we had to transform the actual customer data into a trainable dataset. Therefore, in this study, by leveraging the constructed data quality evaluation indicator system, we introduced a related label generation algorithm. Subsequently, we used the overall score of the single-table quality evaluation as the output indicator, which represents the data quality. Meanwhile, we used the scores of the six primary indicators as the input indicators for the machine learning-based quality evaluation model.

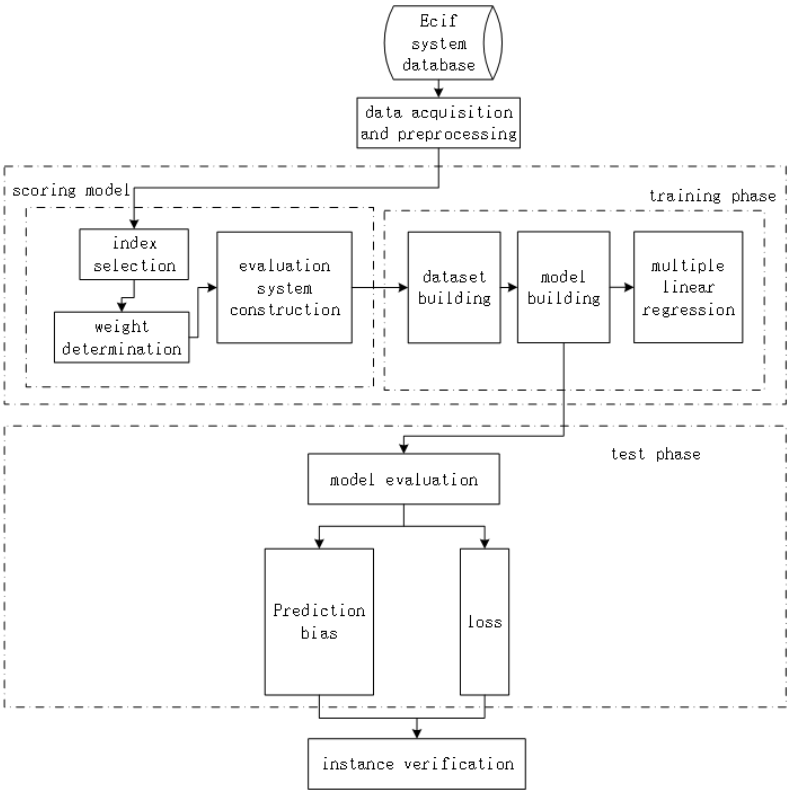


Figure 1. Quality evaluation model based on multiple linear regression.

3.2. Label Generation Algorithm

Suppose there are p primary indicators ($i = 1, 2, \ldots, p$), and each primary indicator i has q_i secondary indicators. A given data table is scored based on the data quality evalua-

tion system established. Specifically, Equation (3) is utilized to generate the corresponding labels for this data table:

$$y = \sum_{i=1}^{p} \alpha_i \sum_{j=1}^{q_i} \beta_j x_j \tag{3}$$

where α_i represents the weight corresponding to each primary indicator, and β_j represents the weight corresponding to the secondary indicator. In the end, each table is associated with a quality evaluation score, which serves as the label for that table.

On the basis of the existing large data table, we randomly sampled different numbers of data entries and generated data subtables. We then applied the label generation algorithm to produce labels y_n for each table. Next, we compiled a trainable assessment table $X = \{X_i\}, X_i = \{x_j\}$ based on each table's primary indicators and their respective final quality evaluation scores. This process led to the formation of a dataset that includes n samples.

To examine the relationships among the dependent variables and between the dependent and independent variables, we conducted a Pearson correlation analysis of the data. Once the correlations among the indicators were determined, we were able to identify and remove the correlated indicators to streamline the model and reduce the training time.

3.3. Feature Selection Based on Shapley Value

Feature selection and hyperparameter tuning are two important steps in every machine learning task. They help improve performance most of the time but have the disadvantage of being expensive in time. The more parameter combinations, or the more precise the selection process, the longer the duration. The usual approach is to combine tuning and feature selection. For feature selection, we adopt a ranking-based selection algorithm. Rank selection involves iteratively removing less important features while retraining the model until convergence is reached. SHAP helps when we perform feature selection using ranking-based algorithms. Instead of using the default variable importances generated by gradient boosting, we choose the best features, e.g., the one with the highest Shapley value.

The Shapley value originates from cooperative game theory and is a distribution method based on contribution. It requires that profits and costs be fairly shared among each agent in an alliance [22]. The Shapley value has numerous applications in machine learning, including data pricing, federated learning, interpretability, reinforcement learning, and feature selection [23].

The equation for the Shapley value of n members is as follows [24]:

$$\phi_i(v) = \sum_{S \subset N, i \in S} \frac{(|S|-1)!(n-|S|)!}{n!} [v(S) - v(S - \{i\})], i = 1, \ldots, n, \tag{4}$$

where $\frac{(|S|-1)!(n-|S|)!}{n!}$ represents the probability of member i joining the alliance $(|S| - i)$. The denominator represents the number of permutations of n members. The numerator represents the number of permutations of the first $(|S| - 1)$ members entering the alliance $(|S| - i)$. Upon joining the alliance, member i acquires a certain value, denoted as $[v(S) - v(S - \{i\})]$. This value represents the contribution that member i makes to the alliance S.

In numerous scenarios involving data mining and machine learning, we frequently encounter challenges, such as the high dimensionality of features [25] and unknown relationships. In this study, we adopted the Powershap algorithm to select and eliminate indicators [26].

The following steps outline the automated feature selection methodology implemented through Powershap, which is constructed based on the foundation of the genetic algorithm:

1. Population initialization: A set of feature subsets is randomly produced, marking the formation of the initial population.

2. Fitness evaluation: A fitness function is implemented to assess the aptness of each entity, employing metrics like accuracy, F1 score, and so forth.
3. Selection process: Individuals exhibiting higher fitness are chosen to serve as the base for the next-generation population as determined by the fitness function assessment results.
4. Crossover procedure: Using a crossover operation, select individuals with superior fitness are randomly amalgamated to generate new offspring.
5. Mutation procedure: To enhance the population's diversity, random alterations are introduced to the new offspring through a mutation operation.
6. Iteration: Steps 2–5 are reiterated until the predetermined stop conditions are satisfied.

3.4. Regression Model Building

Linear regression is a linear model that assumes a linear relationship between an input variable and a single output variable.Specifically, using the linear regression model, the output variable y can be calculated from the linear combination of a set of input variables x, that is, $y = ax + b$. If there are two or more independent variables, such linear regression analysis is called multiple linear regression (MLP). The multiple linear regression model is a relatively simple regression model in machine learning. After using the Shapley value to eliminate model features, the weights calculated based on AHP will no longer be accurate. Therefore, consider using MLP to recalculate the weights.

In the MLP model, it assumes that if the secondary indicators are reasonably set, the quality of the data table is considered only from the dimensions of the primary indicators [27]. If the primary indicator value and the quality of the data table are related linearly, a multivariate regression model can be established [28]: $y = \sum_i^p a_i x_i + b$. The most suitable $\{a_i\}$ can be solved by solving the MLP problem. We employed the MLP solution method, specifically gradient descent, to determine the optimal parameters for the indicators. This was achieved after eliminating irrelevant features from the model.

4. Results

Financial institutions, such as trusts, have a large number of customers. Analyzing the current status of customers is extremely important for enhancing the accuracy of customer marketing and improving the user experience of customer services [29]. This section takes the personal customer table and institutional customer table in the customer information integration system of a trust institution as examples to illustrate the practical application of the data quality evaluation model.

4.1. Quality Evaluation Indicator System

By following the AHP procedure and the selected primary and secondary indicators in Figure 2, the decision makers have to indicate preferences or priority for each decision alternative in terms of how it contributes to each criterion as shown in Table 5.

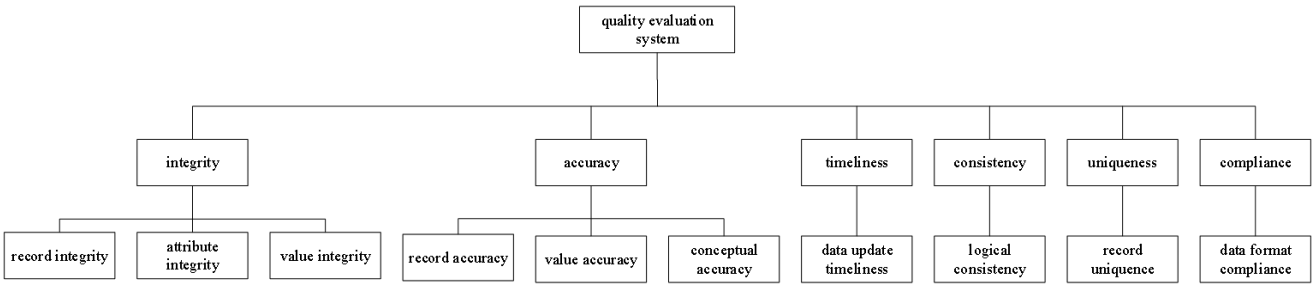


Figure 2. Dimensions of quality evaluation system.

Then, the following can be performed manually or automatically by the AHP software Expert Choice:

1. Synthesizing the pair-wise comparison matrix (example: Table 5).

2. Calculating the priority vector for a criterion such as experience (example: Table 5).
3. Calculating the consistency ratio.
4. Calculating λ_{max}.
5. Calculating the consistency index, CI.
6. Selecting appropriate value of the random consistency ratio from Table 5.
7. Checking the consistency of the pair-wise comparison matrix to check whether the decision maker's comparisons were consistent or not.

Table 5. Average random consistency (RI).

Size of Matrix	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Random consistency	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

w is the eigenvector corresponding to the largest eigenvalue.

Now, we find the consistency index, CI, as follows:

$$CI = \frac{\lambda_{max} - n}{n - 1} = \frac{6.068 - 6}{6 - 1} = 0.0136 \quad (5)$$

Selecting an appropriate value of random consistency ratio RI for a matrix size of five using Table 5, we find $RI = 1.24$. We then calculate the consistency ratio CR as follows:

$$CR = \frac{CI}{RI} = \frac{0.0136}{1.24} = 0.011 \quad (6)$$

As the value of CR is less than 0.1, the judgments are acceptable. Similarly, the pair-wise comparison matrices and priority vectors for the remaining criteria can be found as shown in Tables 6–9, respectively.

Regarding completeness, out of the 264,762 records, including 14,709 data items and 18 attributes, we obtained 67,839 records with null values. This accounted for approximately 25.62% of the data, resulting in a completeness score of 81.70.

Table 6. Quality evaluation system evaluation matrix $A1$.

	$B1$ **Integrity**	$B2$ **Accuracy**	$B3$ **Timeliness**	$B4$ **Consistency**	$B5$ **Uniqueness**	$B6$ **Compliance**	w_i
$B1$ integrity	1	2	1/2	3	3	2	0.229
$B2$ accuracy	1/2	1	1/3	2	2	1	0.133
$B3$ timeliness	2	3	1	4	4	3	0.364
$B4$ consistency	1/3	1/2	1/4	1	1	1	0.084
$B5$ uniqueness	1/3	1/2	1/4	1	1	1	0.084
$B6$ compliance	1/2	1	1/3	1	1	1	0.106
$\lambda_{max} = 6.068, CR = 0.011 < 0.1$							

Table 7. Integrity assessment matrix $A2$.

	$C11$ **Record Integrity**	$C12$ **Attribute Integrity**	$C13$ **Value Integrity**	w_i
C11 record integrity	1	5	5	0.714
C12 attribute integrity	1/5	1	1	0.143
C13 value integrity	1/5	1	1	0.143
$\lambda_{max} = 3.000, CR = 0.000 < 0.1$				

Table 8. Accuracy evaluation matrix *A*3.

	*C*21 **Logical Accuracy**	*C*22 **Value Accuracy**	*C*23 **Conceptual Accuracy**	w_i
*C*21 logical accuracy	1	2	1/3	0.238
*C*22 value accuracy	1/2	1	1/4	0.137
*C*23 conceptual accuracy	3	4	1	0.625
$\lambda_{max} = 3.018, CR = 0.000 < 0.1$				

Table 9. Weights of quality bid evaluation indicators.

Level 1 Indicators			**Level 2 Indicators**		
indicator number	indicator name	weights (μ_m)	indicator number	indicator name	weights (μ_n)
B1	integrity	0.229	C11	record integrity	0.714
			C12	attribute integrity	0.143
			C13	value integrity	0.143
B2	accuracy	0.133	C21	logical accuracy	0.238
			C22	value accuracy	0.137
			C23	conceptual accuracy	0.625
B3	timeliness	0.364	C31	data update timeliness	1
B4	consistency	0.084	C41	logical consistency	1
B5	uniqueness	0.084	C51	record uniqueness	1
B6	compliance	0.106	C61	format compliance	1

Concerning accuracy, among the 14,709 data records, we did not have any instances of data being updated or created later than the current time. Additionally, we had 5191 records with expired certificates. Roughly 35.29% of the data exhibited accuracy issues, resulting in an accuracy score of 91.58.

In terms of timeliness, we did not identify any problems within the 14,709 data records analyzed, warranting a perfect timeliness score of 100.

In terms of consistency, out of the 14,709 data records, 13,869 records belonging to the same information subject within the dataset did not meet the criteria for logical accuracy. This suggested that approximately 94.29% of the data exhibited consistency issues, resulting in a consistency score of 5.71.

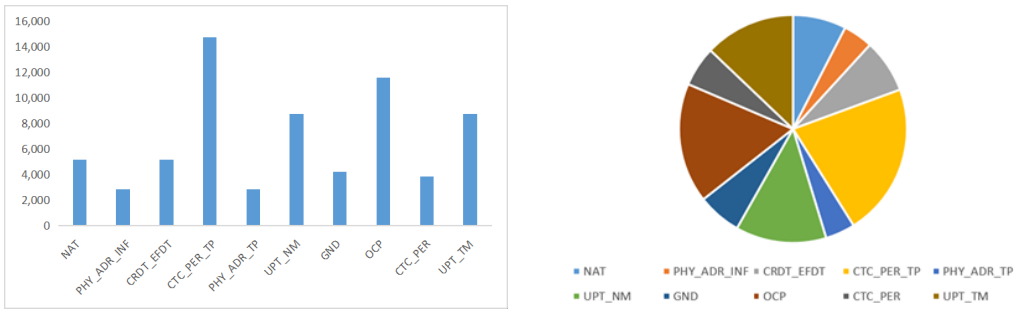
Regarding uniqueness, we did not find any instances of noncompliant record uniqueness among the 14,709 data records, resulting in a uniqueness score of 100.

In terms of compliance, among the 14,709 data records, 169 records did not conform to the requirement of 18-digit identity card numbers. Further scrutiny of the birth year, month, day, and last digit did not reveal any records that violated the year and month requirements. However, 1380 records did not comply with the day requirement. Additionally, for contact information, 4908 records had noncompliant mobile phone numbers. Overall, the number of noncompliant data items reached 6457, accounting for approximately 21.95% of the total data. Consequently, the compliance score was calculated as 58.43.

Considering the six dimensions of data quality and the data quality evaluation model, the individual quality evaluation score for the consignment customer form was determined to be 82.39 as shown in Table 10 and Figure 3.

Table 10. Individual customer completeness statistical results.

Field Chinese Name	Field English Name	Is Empty	Null Ratio	Field Chinese Name	Field English Name	Is Empty	Null Ratio
name	CST_NM	0	0	type of certificate	CRDT_TP	0	0
ID number	CRDT_NO	0	0	gender	GND	4225	28.72%
nationality	NAT	5145	34.98%	profession	OCP	11,549	78.52%
address	PHY_ADR_INF	2859	19.44%	contact information	CTC_PER	3850	26.17%
certificate validity period	CRDT_EFDT	5156	35.05%	customer number	CST_ID	0	0
contact type	CTC_PER_TP	14,709	1	system source	SYS_SRC	0	0
address type	PHY_ADR_TP	2858	19.43%	sales channels	SALE_SRC	0	0
updater	UPT_NM	8744	59.45%	update time	UPT_TM	8744	59.45%
founder	CRT_NM	0	0	creation time	CRT_TM	0	0



(**a**) The number of incomplete fields (**b**) Incomplete proportion of each field

Figure 3. Integrity statistics results.

4.2. Multiple Linear Regression Quality Evaluation Model

We created the evaluation dataset using the label generation algorithm and the quality evaluation indicator system. The descriptive statistical results of various indicators of the trust agency's individual customer table and institutional customer table are shown in Tables 11 and 12.

Table 11. Descriptive statistics of indicators in the personal customer table.

Index	Maximum Value	Minimum Value	Average	Standard Deviation
integrity	87.59	74.4	80.9	0.25
accuracy	98.3	80.26	91.56	0.38
timeliness	100	100	100	0
consistency	9.09	0	1.15	0.23
uniqueness	100	52.61	99.94	1.26
Compliance	74.12	6.67	67.53	1.71

Table 12. Descriptive statistical table of indicators in the institutional client table.

Index	Maximum Value	Minimum Value	Average	Standard Deviation
integrity	84.75	78.81	80.12	0.3
accuracy	97.83	82.1	92.33	0.28
timeliness	100	100	100	0
consistency	9.3	1.5	3.85	0.29
uniqueness	100	89.66	98.1	0.86
Compliance	50.66	11.5	43.21	2.03

To examine the presence of correlations among dependent variables and between the dependent and independent variables, we performed a Pearson correlation analysis of the data. The results were visualized through a heatmap as depicted in Figure 4, showcasing the correlations between the variables. In the heatmap, black represents a negative correlation, white represents a positive correlation, and the diagonal line represents the correlation of each independent variable with itself, resulting in a correlation coefficient of 1.

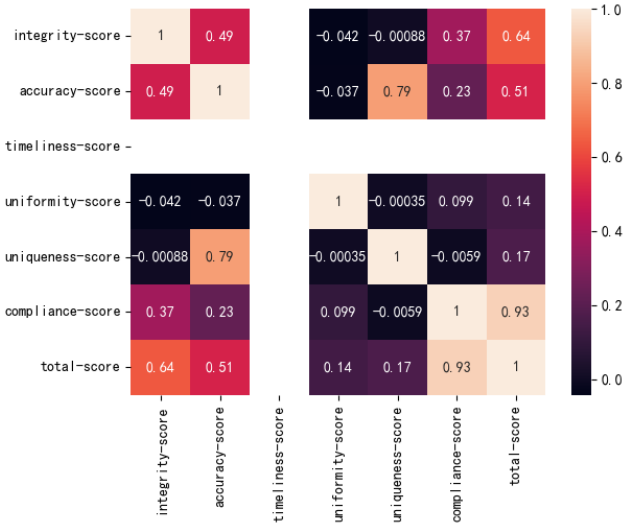
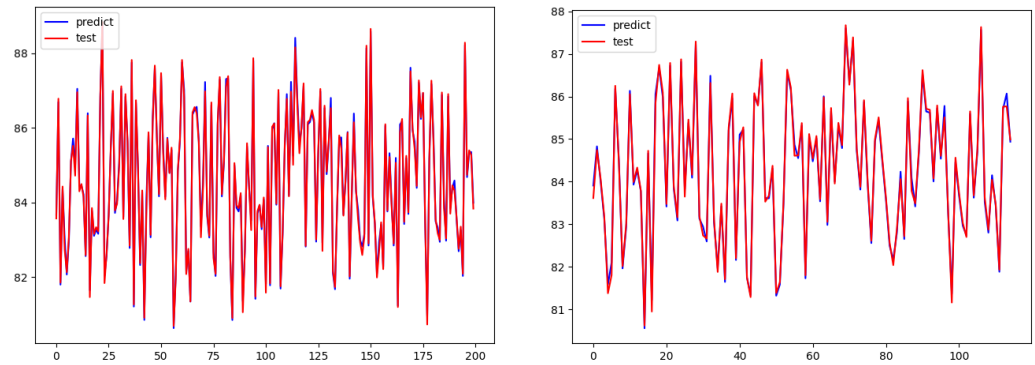


Figure 4. Heat map of correlation between variables.

We observed that the correlation between the various indicator variables was not notably high, suggesting their independence from each other and the absence of multicollinearity. Moreover, the correlation between the independent variables and dependent variables was quite strong. Among these correlations, the compliance_score exhibited the highest correlation, with a coefficient of 0.93. This result indicated that compliance, as one of the first-layer indicators, exerted the most significant influence on the final quality score. Additionally, the integrityaccuracy_score also demonstrated notable correlations with the quality score, with coefficients of 0.64 and 0.59, respectively.

The modeling approach for the MLP model in this study had the following steps: (1) The indicator data and quality score data were normalized. (2) The dataset was divided into a training set (70% of the data) and a testing set (remaining 30% of the data). (3) The MLP model was constructed using the gradient descent algorithm to derive the optimal weight values for the model. The experimental results are shown in Figure 5 and 6.



(**a**) Personal customer data (**b**) Institutional client data

Figure 5. True value and predicted value of quality evaluation score.

The Figure 5 shows the results of the real value and predicted value of the test set. Among them, the blue is the predicted value, and the red is the real value. It is evident from the figure that a discernible variance exists between the predicted and actual values. The magnitude of this difference is relatively small, however, which suggests a high level of accuracy in the prediction model.

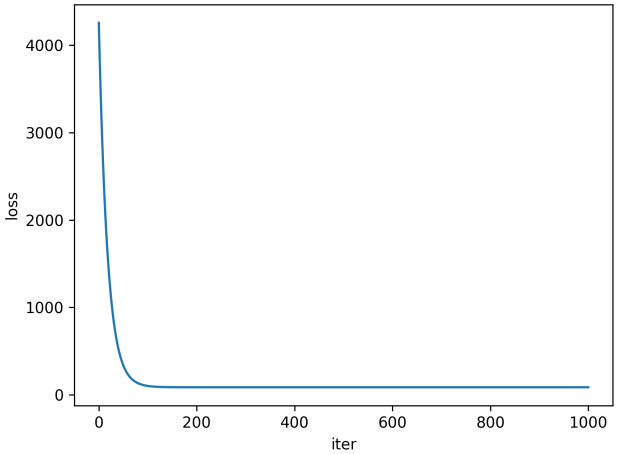


Figure 6. Test set loss function graph.

In real-world applications, each sample represents a distinct customer table. Tables 4–7 demonstrate the evaluation of the consignment personal customer information table and the general personal customer information table from the trust institution using two different methods: the conventional scoring technique and a machine learning-based method. The time invested and the quality assessment scores obtained from the AHP are approximate representations of the outcomes achieved by the manual scoring method currently employed by the trust.

According to the incremental experimental results, we observed that the time required by both methods increased as the sample size grew. When the number of customer tables doubled, the time spent on the traditional method also doubled, whereas the time expenditure for the machine learning-based approach experienced only a slight increase. Moreover, the quality assessment scores generated through the MLP method were relatively higher than the traditional quality evaluation techniques. By employing Powershap for feature selection, we excluded redundant data, thus reducing the time expenditure and improving score precision(Table 13).

Table 13. The results for samples of different orders of magnitude were calculated based on the AHP and the method proposed in this paper.

Table Name	Number of Samples	AHP Time Spent	Quality Review Score	Machine Learning Time Spent	Quality Review Score
MO_CI_PER_CONSIGNMENT	10,000	600.02 s	81.39	613.89 s	83.55
	20,000	1192.28 s	83.21	650.25 s	85.89
	30,000	1532.43 s	80.91	690.11 s	83.22
MO_CI_PER_CUSTOMER	10,000	576.22 s	84.43	623.14 s	85.99
	20,000	1312.34 s	79.98	652.25 s	82.01
	30,000	1701.36 s	80.12	661.21 s	80.56

4.3. Feature Selection Based on Shapley Value

We initially split the data into a 3:7 ratio for testing and training purposes. Subsequently, we categorized the data into six classes, representing the six primary indicators.

Then, we constructed a Powershap function, utilizing the chosen classification model, which, in this study, was the logistic regression CV algorithm.

When employing the logistic regression CV algorithm model for 200 iterations, we observed that significant features accounted for five-sixths of the total (Figure 7). By considering impact scores and other statistical results, the timeliness factor could be excluded from the primary indicators. This exclusion not only reduced the required effort, but also improved the precision in determining the weight values for subsequent analysis.

	impact	p_value	...	power_0.01_alpha	0.99_power_its_req
compliance_score	2.755706	0.2	...	0.0	0.0
integrity_score	0.851455	0.6	...	0.0	0.0
random_uniform_feature	0.824305	0.6	...	0.0	0.0
accuracy_score	0.734240	0.7	...	0.0	0.0
uniqueness_score	0.727508	0.4	...	0.0	0.0
uniformity_score	0.239986	1.0	...	0.0	0.0
timeliness_score	0.000000	1.0	...	0.0	0.0

[7 rows x 5 columns]

Figure 7. The importance of each feature.

5. Conclusions

This article organizes and analyzes the existing quality evaluation indicators, and comprehensively uses completeness, accuracy, timeliness, consistency, uniqueness and compliance to establish a quality evaluation indicator system for trusts and other financial institutions. Preliminary weight determination is made based on AHP. Since there are many indicators, the Shapley value is introduced for feature selection. A multiple linear regression model is used to determine new weights for the screened indicators. The model is verified using customer data and institutional data from a trust company's customer information integration system.

Through data quality scoring, trust data can be graded for quality, thereby improving data quality and promoting the transfer and transformation of data results from trusts and other financial institutions. Using machine learning regression algorithms to evaluate data quality has the advantages of quantifiable evaluation results, more intuitive results, and higher accuracy. In addition, in the face of massive customer data, the intelligent data quality evaluation model based on machine learning is faster and more convenient, and the evaluation results are more accurate and objective.

Therefore, the data quality evaluation model based on the regression algorithm model can not only be used to evaluate the quality of trust customer data but can also be used in other industries and government departments to evaluate the quality of other types of data, including their own, as well as the benefits and value that the data can bring.

Author Contributions: Conceptualization, M.L. and Y.Y.; methodology, M.L.; software, M.L.; validation, M.L., J.L. and Y.Y.; formal analysis, M.L.; investigation, J.L.; resources, Y.Y.; data curation, M.L.; writing—original draft preparation, M.L.; writing—review and editing, J.L.; visualization, M.L.; supervision, J.L.; project administration, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Weber, K.; Otto, B.; Osterle, H. One Size Does Not Fit All—A Contingency Approach to Data Governance. *J. Data Inf. Qual.* **2009**, *1*, 4.
2. Begg, C.; Cairat, T. Exploring the SME quandary:data governance in practise in the small to medium-sized enterprise sector. *Electron. J. Inf. Syst. Eval.* **2012**, *15*, 3–13.

3. Newman, D.; Logan, D. Governance is an essential building block for enterprise information management. *Gart. Res. Stamford* **2006**, *13*, 4.
4. Niemi, E. Designing a data governance framework. In Proceedings of the IRIS Conference, Turku, Finland, 16–19 August 2011.
5. Karkošková, S. Data governance model to enhance data quality in financial institutions. *Inf. Syst. Manag.* **2023**, *40*, 90–110.
6. Mcgilvray, D. *Executing Data Quality Projects Ten Steps to Quality Data and Trusted Information*; Elsevier: Amsterdam, The Netherlands, 2008; Volume 12, pp. 2–4.
7. Omara, E.; Said, T.E.; Mousa, M. Employing neural networks for assessment of data quality with emphasis on data completeness. *Int. J. Artif. Intell. Mach. Learn.* **2011**, *11*, 21.
8. Peltier, J.W.; Zahay, D.; Lehmann, D.R. Organizational learning and CRM success: A model for linking organizational practices, customer data quality, and performance. *J. Interact. Mark.* **2013**, *27*, 1–13.
9. Taleb, I.; Kassabi, H.; Serhani, M.A.; Dssouli, R.; Bouhaddioui, C. Big Data Quality: A Quality Dimensions Evaluation. Ubiquitous Intelligence and Computing. In Proceedings of the Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, Toulouse, France, 18–21 July 2016.
10. Juddoo, S. Overview of data quality challenges in the context of Big Data. In Proceedings of the 2015 International Conference on Computing, Communication and Security (ICCCS), Pointe aux Piments, Mauritius, 4–5 December 2015; pp. 1–9.
11. Madhikermi, M.; Kubler, S.; Robert, J.; Buda, A.; Främling, K. Data quality assessment of maintenance reporting procedures. *Expert Syst. Appl.* **2016**, *63*, 145–164.
12. Mashoufi, M.; Ayatollahi, H.; Khorasani-Zavareh, D.; Talebi Azad Boni, T. Data quality in health care: Main concepts and assessment methodologies. *Methods Inf. Med.* **2023**, *62*, 5–18.
13. Uzoka, F.M. AHP-based system for strategic evaluation of financial information. *Inf. Knowl. Syst. Manag.* **2005**, *5*, 49–61.
14. Khan, A.W.; Khan, M.U.; Khan, J.A.; Ahmad, A.; Khan, K.; Zamir, M.; Kim, W.; Ijaz, M.F. Analyzing and evaluating critical challenges and practices for software vendor organizations to secure big data on cloud computing: An AHP-based systematic approach. *IEEE Access* **2021**, *9*, 107309–107332.
15. Alam, M.K. A systematic qualitative case study: Questions, data collection, NVivo analysis and saturation. *Qual. Res. Organ. Manag. Int. J.* **2021**, *16*, 1–31.
16. Gomes, V.C.F.; Queiroz, G.R.; Ferreira, K.R. An overview of platforms for big earth observation data management and analysis. *Remote Sens.* **2020**, *12*, 1253.
17. Liu, Y.; Wang, Y.; Xue, F.; Chen, J. A hybrid approach for supplier selection based on quality management system evaluation and grey relational analysis. *J. Intell. Fuzzy Syst.* **2021**, *41*, 1149–1159.
18. Malik, S.; Tahir, M.; Sardaraz, M.; Alourani, A. A resource utilization prediction model for cloud data centers using evolutionary algorithms and machine learning techniques. *Appl. Sci.* **2022**, *12*, 2160.
19. Titus, B.D.; Brown, K.; Helmisaari, H.S.; Vanguelova, E.; Stupak, I.; Evans, A.; Clarke, N.; Guidi, C.; Bruckman, V.J.; Varnagiryte-Kabasinskiene, I.; et al. Sustainable forest biomass: A review of current residue harvesting guidelines. *Energy Sustain. Soc.* **2021**, *11*, 10.
20. Hou, Y.; Biljecki, F. A comprehensive framework for evaluating the quality of street view imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 103094.
21. Sun, H.; Chen, Z. Interval neutrosophic hesitant fuzzy AHP method based on combined weights. *J. Intell. Fuzzy Syst.* **2021**, *41*, 8015–8028.
22. Wang, Y.; Liu, Z.; Cai, C.; Xue, L.; Ma, Y.; Shen, H.; Chen, X.; Liu, L. Research on the optimization method of integrated energy system operation with multi-subject game. *Energy* **2022**, *21*, 123305.
23. Liu, J.; Huang, J.; Zhou, Y.; Li, X.; Ji, S.; Xiong, H.; Dou, D. From distributed machine learning to federated learning: A survey. *Knowl. Inf. Syst.* **2022**, *64*, 885–917.
24. Chen, H.; Covert, I.C.; Lundberg, S.M.; Lee, S.I. Algorithms to estimate Shapley value feature attributions. *Nat. Mach. Intell.* **2023**, *5*, 590–601.
25. Liu, L.; Li, Z.; Yang, J. An emergency plan evaluation model based on combined DEA and TOPSIS methods. *J. Clean. Prod.* **2021**, *315*, 62.
26. Kitiyodom, P.; Chindapa, P. Development of an emergency response plan assessment model for hazardous chemical accidents in Thailand. *J. Loss Prev. Process. Ind.* **2021**, *70*, 307.
27. Wen, C.; Yang, J.; Gan, L.; Pan, Y. Big data driven Internet of Things for credit evaluation and early warning in finance. *Future Gener. Comput. Syst.* **2021**, *34*, 295–307.
28. Liapis, C.M.; Kotsiantis, S. Energy Balance Forecasting: An Extensive Multivariate Regression Models Comparison. In Proceedings of the 12th Hellenic Conference on Artificial Intelligence, Corfu, Greece, 7–9 September 2022; pp. 1–7.
29. Tiwari, P. Bank affection and customer retention: An empirical investigation of customer trust, satisfaction, loyalty. *SN Bus. Econ.* **2022**, *2*, 54.