sustainability

MDPI

Article

Research on Enterprise Digital-Level Classification Based on XGBoost Model

Qiuxia Ren * and Jigan Wang

School of Business, Hohai University, Nanjing 211100, China
* Correspondence: rqx_1021@hhu.edu.cn

Abstract: Digital knowledge and information have become significant production variables that have permeated all aspects of life and play a leading and supporting role in the growth of the real economy as the digital economy has developed. Through field research and web research, this study identifies digital-economy-related enterprises as the survey object; summarizes the fundamental information for these enterprises, their level of digitization, and the dilemma and demands of digital-level advancement; and generates survey data for 1936 enterprises. On the basis of these data, this study extracts the elements that influence the improvement of the enterprises' digital level, applies statistical knowledge and machine learning techniques, and derives an enterprise digitization level index system and associated index score for enterprise digitization level. The experimental results indicate that the region, the time of establishment, the nature of ownership, the number of employees, R&D investment, being a national high-tech enterprise, and the establishment of digital transformation management departments have major effects. The AUC value of the XGBoost model modeled using all feature variables has achieved certain results, and the five assessment indices of the model have been enhanced to varying degrees, with the AUC being 0.9263.

Keywords: digital economy; enterprise digital level; XGBoost model

check for updates

Citation: Ren, Q.; Wang, J. Research on Enterprise Digital-Level Classification Based on XGBoost Model. *Sustainability* **2023**, *15*, 2699. https://doi.org/10.3390/su15032699

Academic Editors: Feng Liu, Peiwan Wang and Jiayin Qi

Received: 21 December 2022
Revised: 28 January 2023
Accepted: 29 January 2023
Published: 2 February 2023

1. Introduction

Presently, the digital economy has become the direction and focus of development for the world's leading nations. The global economy has transitioned from the industrial era to the digital era. The new format and model enabled by digital technology have invigorated the growth of the global economy. The digitization of businesses has started to garner broad interest in worldwide media and academia. China has also implemented a number of macro-level plans and initiatives in the realm of the digital economy. China's digital economy will reach CNY 39.2 trillion in 2020, accounting for 38.6% of the country's gross domestic product, with apparent late-mover benefits. In recent years, the state has placed a high priority on the development of the digital economy, incorporated data into the category of production factors, emphasized the acceleration of the cultivation of the data factor market, enhanced the income distribution mechanism related to data factors, promoted digital transformation, and achieved the deep integration of the digital economy and the real economy. Development has reached the pinnacle of the national strategy for medium- to long-term development.

Gupta pointed out that in the Internet era, digital capabilities have become the organizations' key resources and competencies [1]. Digitization is the transition from a single communication technology to an integrated digital technology, which emerged from informatization and internetization. The understanding of the digital transformation can be summed up as follows: the integrated innovation of new generation digital technologies, such as big data, cloud computing, the Internet of Things, and artificial intelligence, which are deeply integrated with enterprises, regards data as a new factor of production and then promotes the innovation of production organization and business models, and the resource is difficult to imitate and replace [2]. In the wave of development of the digital economy, enterprises are not only the main source of innovation but also the vehicle for the integration and use of digital technologies. Moreover, enterprises can obtain a competitive edge and distinctive operations through the improvement of their digital level. Its essence is to increase the gap with competitors, strengthen consumer loyalty, and enhance corporate performance. Therefore, promoting and achieving the improvement of company digitization levels is crucial to the growth of the digital economy. This study's primary objective is to determine how to quantify and categorize the extent of enterprise digitalization.

The research on the digitalization levels of enterprises focuses mostly on the strategic level, and some scholars believe that digital transformation is dominated by technology to build a new digital economy system, and the overall efficiency of enterprises will be improved through the application of digital technology [3]. Additionally, some academics argue that this represents a shift in the ideological value perspective [4]. Traditional companies will modify their business models by shifting their focus from customer value to digital technology. Nonetheless, some academics believe that digitalization is intimately linked to an enterprise's adopted strategy [5]. Few scholars have built an index system to quantify the level of enterprise digitization, and the classification of enterprise digitization levels is rarely discussed in the current literature.

Through field research and web research, this study identifies digital-economy-related enterprises as the survey object; compiles the fundamental information for these enterprises, their level of digitization, and the dilemma and demands of digital level improvement; and generates survey data for 1936 enterprises. Based on these statistics, this report identifies the characteristics that have an impact on the enterprises' digital advancement. Using statistical knowledge and a machine learning model, this paper constructs an index system that affects the digital level of the enterprise and categorizes the enterprises' digital transformation levels based on the enterprises' basic information, such as their regions, times of establishment, and the nature of the enterprises. For the independence test of categorical variables, the multi-classification logistic regression method is further used on the basis of contingency table analysis to further explore the impact of the basic information for the enterprises regarding the improvement of their digitization level. This study can provide theoretical guidance for the future advancement of the digital levels of firms and expand the theory of digital economic development by combining its findings with those of earlier research. Simultaneously, the machine learning research method is integrated into the current enterprise management practice based on the rapid expansion of the digital economy at the macro level, promoting industrial upgrading and transformation.

2. Related Works

With the wave of digitization sweeping the world, research on the level of enterprise digitization continues to emerge, but it is still in its exploratory phase, and there is a dearth of integrated research on the impact mechanism and stage of digital-level improvement. Sachs believed that an enterprise's digital transformation necessitates the establishment of a new production and management system [6]. Industrial digitization is the modernization of the agriculture, manufacturing, and service industries by digital technology, and it is the goal of the digital transformation of enterprises. Sebastian et al. and Bharadwaj et al. believe that digital transformation should include changes in the way companies work and organize and their overall business models as caused by the combination of different technologies triggered by digital technologies [7,8]. Enterprises must employ 5G, big data, cloud computing, the Internet of Things, and other digital technologies to the purchase, production, and circulation of raw materials, as well as other links. It is the implementation of industrial digitization and represents the value of data as a production element.

External variables that influence digital transformation include primarily the growth and diffusion of digital technology, the intensification of the competitive environment, and the evolution of user demand. Matt et al. noted that the growth of digital technology can expand the limits of organizations and facilitate the shift of their strategic focus from product-centered to customer-centered [9]. In turn, the improvement of information sharing among enterprises can promote the adoption of big data analysis technology, form a virtuous circle, and enhance supply chain performance and organizational performance [10]. Bellalouna studied the key role of augmented reality (AR) in the field of manufacturing digitization and demonstrated the application of AR in digital product configuration management and production planning assistance with case studies [11]. Li et al. believe that the continuous digital technology revolution can improve the coordination between business strategy and digital technology, hence fostering corporate growth [12]. From the perspective of global value chains, Wu et al. assert that the digitization of the manufacturing industry will improve the comparative advantage and competitiveness of the country in the global value chain [13]. In a competitive context, firms launch digital innovation mechanisms, and digital innovation can also alter the competitive climate in which enterprises operate [14]. For instance, the use of remote information technology by the Italian vehicle insurance business Unipol signifies a strategic move in the market toward data-driven services, such as insurance rates based on driving behaviors [15]. Abrell et al. analyzed three B2B organizations dealing with digital innovation and determined the various roles of customers and users in guiding digital innovation [16].

Digital transformation's primary function is to foster product and service innovation, process innovation, and business model innovation, ultimately improving operational efficiency and organizational performance. Bharadwaj et al. indicated that the digitization of business infrastructure tightens the connections between products, processes, and services, which can significantly enhance products and services and increase company networks [17]. In the context of digital innovation, innovation subjects are more diversified, innovation is more democratic, the distinction between innovation processes and outcomes is shattered, product boundaries become mobile, and innovation activities continue during user usage [18]. Boersma et al. believed that the application of Internet of Things technology has spawned a new business model and enhanced the profitability of the enterprise supply chain [19]. Gnimpieba et al. constructed a digital architecture platform utilizing cloud computing technology and Internet of Things technology [20]. Warner et al. explored how existing enterprises in traditional industries can build dynamic capabilities for digital transformation, simplify operations, or create new business models to enhance customer experience [21].

Reviewing the literature in this field, many researchers utilize classic research methodologies, such as theoretical research, case studies, quantitative analysis, etc. to analyze the digital levels of organizations. Hazen et al. analyzed the impact of digital technology on supply chain sustainability from eight theoretical perspectives and concluded that, as digital technology continues to mature and be applied, it can continue to provide supply chain competitive advantages [22]. Wang et al. summarized the pertinent literature on the application of big data in supply chain management, examined the potential application scenarios of big data technology in the future, and emphasized the significance of big data technology in enterprise supply chain management [23]. Hess et al. investigated the role of CIOs (chief information officers) and other senior managers in responding to the risks and opportunities of digital transformation and developed a digital transformation strategy guidance matrix based on the successful digital transformation of three German media companies [24]. Wei Zhang et al. employed Hamilton–Jacobi–Bellman equation to quantify the industrial digital collaboration mechanism and optimal strategy within the setting of the digital economy [24].

At present, case studies are mainly used to study the transformation and upgrading of digital enterprises and value creation, which is conducive to exploring the mechanism of digital innovation and deeply dissecting the differences between digital innovation and traditional innovation models from an internal perspective. However, the empirical research rooted in the context of the digital economy is insufficient, including the measurement methods and tool selection of digital innovation, which not only shows that there are still a lot of gaps in the research field of enterprise digital level but also shows that the research on enterprises' digital levels is still not systematic. With the understanding of the dimensions of digital innovation, future research can explore the research methods of scale or large sample empirical analysis for further verification [25]. As a prevalent research approach in the field of social science, machine learning intends to employ information technology to improve the quantitative analytical level of social science research in order to solve multidimensional, multivariable, and dynamically complicated problems.

To sum up, the research on the related and derivative domains of the digital economy is developing increasingly sufficiently to adapt to the evolution of the times. However, for the measurement of enterprises' digital levels, the existing research mainly focuses on the impact of the improvement of the enterprise digital level on human resources, financial management, incentive mechanism optimization, value chain cooperation, and so on. At the same time, the construction of the influencing factors of the digital transformation of enterprises and the calculation of the index weight are still in a relatively broad field and primarily discuss the influence of external factors, such as technological development and penetration, intensified competitive environment, changes in user needs, etc., rather than the enterprise itself. This study is primarily based on the theory of statistics and machine learning, starting from the microscopic perspective, and examines some of the conditions of the enterprise 's own information as the influencing factors; the study also constructs an index system and assesses the significance of the characteristics based on the index score. Concurrently, it conducts a classification study on the digital level of enterprises and compensates for the deficiencies in the existing research and provides a valuable reference for expanding the research on the enterprise digitization level and practices for enhancing the enterprise digitization level.

3. Data and Methodology

3.1. Experimental Data

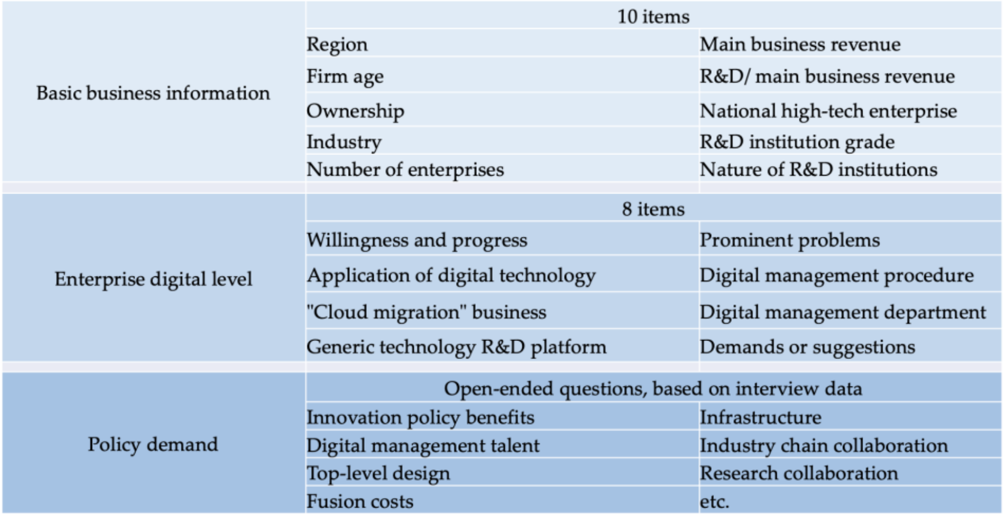
Due to the lack of a unified and precise definition of the digital economy, there are significant regional variations in the assessment of the digital economy [26]. In this study, the data for enterprise digital level classification are obtained by questionnaire survey. The questionnaire consists of three sections: enterprise fundamentals, enterprise digital level state, and digital transformation policy requirements. Among them, the following are found:

(1) Fundamental information. Its survey items include the city where the enterprise is located, the time of establishment, the nature of ownership, the number of employees, and the proportion of R&D investment relative to the business's primary revenue over the preceding 3 years.

(2) Digital level status. It mainly investigates related enterprises from eight aspects—the willingness and promotion of enterprise digital transformation, the development of digital transformation, the driving factors of digital transformation, the application of digital technology, the development of cloud business, the existence and evaluation of industry-wide technology platforms, the outstanding problems of digital transformation, and the means of promoting digital transformation.

(3) The necessity for digital transformation policies. Its primary purpose is to listen to the demands of research organizations and provide decision-making guidance for the formulation of financial support programs from the enterprise's policy-related focus on digital transformation, the enterprise's demands, or ideas for promoting digital transformation to the government.

Each section specifically sets up 1–2 open-ended questions for the investigator to obtain more detailed information. In the context of the topic, these three elements are mostly considered: First, the enterprise's basic information, such as the number of employees, the nature of ownership, and its region, is the enterprise attribute variable, which must be classified and identified. Second, the enterprise's R&D and external technology acquisition are enterprise development variables, reflecting the enterprise's R&D investment and the technical source of digital transformation. Third, the policy requirements of digital transformation, which are mostly dependent on the subjective cognition of businesses, require that we synthesize the viewpoints of businesses and prepare for policy design. In terms of option design, this study starts from both the supply and demand sides. On the demand side, we analyze the driving factors of the improvement of the digitalization levels of enterprises. On the supply side, data factors enable enterprises to promote the intelligent transformation of manufacturing technology mainly by affecting the production organization, resource allocation, and supply modes of enterprises. The related factors of questionnaire design are shown in Figure 1.



Basic business information	10 items	
	Region	Main business revenue
	Firm age	R&D/ main business revenue
	Ownership	National high-tech enterprise
	Industry	R&D institution grade
	Number of enterprises	Nature of R&D institutions
Enterprise digital level	8 items	
	Willingness and progress	Prominent problems
	Application of digital technology	Digital management procedure
	"Cloud migration" business	Digital management department
	Generic technology R&D platform	Demands or suggestions
Policy demand	Open-ended questions, based on interview data	
	Innovation policy benefits	Infrastructure
	Digital management talent	Industry chain collaboration
	Top-level design	Research collaboration
	Fusion costs	etc.

Figure 1. Enterprise digital-level questionnaire design-related factors.

This questionnaire was distributed via the Internet. Using a stratified sampling strategy, representative or typical businesses were studied and online forms were filled in. The distribution period was between 11 June 2022 and 14 July 2022. The respondents included digital-economy-related firms encompassing various sorts of enterprises with diverse ownership structures, establishment dates, employee counts, R&D investment proportions, etc. On the date of questionnaire recovery, a total of 1936 valid samples were received.

Through the establishment of a cross-contingency table for the questionnaire data for 1936 enterprises, the chi-square independence test was carried out. The seven basic information fields of the questionnaire/responses regarding the enterprises (enterprise region, enterprise age, ownership, number of employees, R&D investment, digital management department, and national high-tech enterprises) have a significant impact on the improvement of the digital level of the enterprise. Figure 2 depicts the elements that affect the enterprise's digital maturity level.

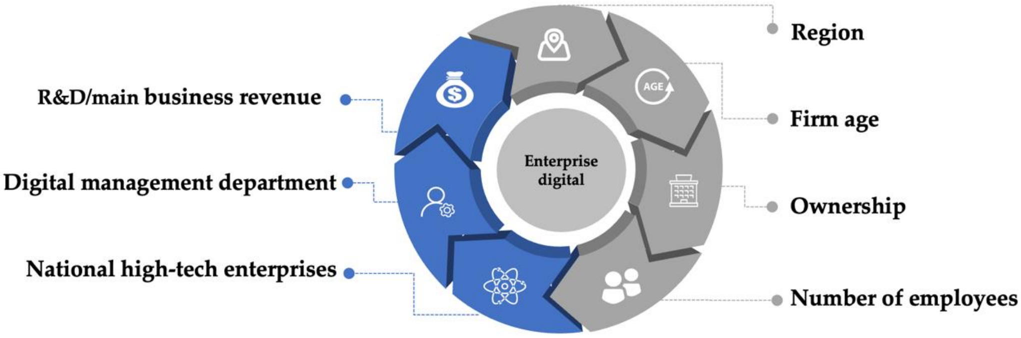

Region
R&D/main business revenue
AGE
Firm age
Enterprise digital
Digital management department
Ownership
National high-tech enterprises
Number of employees

Figure 2. Influencing factors for enterprise digitalization level.

3.2. XGBoost Classification Algorithm

The base learner of the XGBoost classification algorithm is a decision tree model, and the classification features are selected according to the CART algorithm. Different from the random forest model based on Bagging, XGBoost is based on Gradient Boosting.

The XGBoost model is defined as:

$$\hat{y}_i = F_k(x_i) = F_{k-1}(x_i) + f_k(x_i) \tag{1}$$

In Equation (1), $f_k(x)$ represents the kth decision-tree-based learner.

The objective function of XGBoost is defined as follows:

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

The objective function Obj in Formula (2) consists of two items: the former is the loss function; the latter is a regularization term, which is mainly used to control overfitting. The definition of the regularization term is as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel \omega \parallel^2 \tag{3}$$

The optimization goal of the XGBoost model is to find the optimal $f(x_i)$ so that the objective function Obj is minimized. In order to find the optimal $f(x_i)$, XGBoost adopts the method of objective function approximation. For this purpose, Formula (2) is rewritten as follows:

$$Obj^{(s)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(s-1)} + f_s(x_i)\right) + \Omega(f_s) \tag{4}$$

Formula (4) is further subjected to second-order Taylor expansion:

$$Obj^{(s)} \cong \sum_{i=1}^{n} \left[L\left(y_i \hat{y}^{(s-1)}\right) + g_i f_s(x_i) + \frac{1}{2} h_i f_s^2(x_i) \right] + \Omega(f_s) \tag{5}$$

In Formula (5), g_i, h_i are:

$$g_i = \frac{\partial L(y_i, \hat{y}^{(s-1)})}{\partial \hat{y}^{(s-1)}} \tag{6}$$

$$h_i = \frac{\partial^2 L(y_i, \hat{y}^{(s-1)})}{\partial \hat{y}^{(s-1)}} \tag{7}$$

Since the constant term does not affect the optimization results of the model, the Formula (5) is further simplified. Removing the constant term, Equation (5) is converted to:

$$\begin{aligned} Obj^{(s)} &= \sum_{i=1}^{n} \left[g_i f_s(x_i) + \tfrac{1}{2} h_i f_s^2(x_i) \right] + \gamma T + \tfrac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \tfrac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \end{aligned} \tag{8}$$

Through such transformation, we find that Formula (8) is a unary quadratic function with independent variable ω_j and dependent variable $Obj^{(s)}$. According to the formula for finding the most value of unary quadratic function, the optimal ω_j^* of leaf node j is:

$$\omega_j^* = \frac{\sum_{i \in I_j} g_i}{-2 \times \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right)} = \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{9}$$

At this point, the optimal objective function value $Obj^{(s)}$ is:

$$Obj^{(\mathrm{s})} = -\frac{1}{2} \sum_{j=1}^{\mathrm{T}} \frac{\left(\sum_{i \in I_j} g_{\mathrm{i}}\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma \mathrm{T} \tag{10}$$

In order to simplify the expression, in Equations (9) and (10), G_j denotes $\sum_{i \in I_j} g_i$, and H_j denotes $\sum_{i \in I_j} h_i$, then:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{11}$$

$$Obj^{(s)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{12}$$

Equation (12) is the scoring function of a decision-tree-based learner with training. In theory, for each training process, the optimal decision tree model can be obtained by calculating the evaluation scores for all candidate decision tree models. However, in practical problems, the number of candidate decision trees is infinite. Therefore, in order to solve this problem, the answer given by XGBoost is to use a greedy algorithm:

$$Obj_j = -\frac{1}{2} \frac{G_j^2}{H_j + \lambda} + \gamma \tag{13}$$

Because there is only one node at this time, the contribution at this time is γ. After the node is split into two child nodes, the objective function contribution of the two child nodes becomes:

$$Obj_s = -\frac{1}{2} \left(\frac{G_{jL}^2}{H_{jL} + \lambda} + \frac{G_{jR}^2}{H_{jR} + \lambda} \right) + 2\gamma \tag{14}$$

The objective function changes to:

$$Obj_{split}^{(j)} = Obj_j - Obj_s = \frac{1}{2} \left(\frac{G_{jL}^2}{H_{jL} + \lambda} + \frac{G_{jR}^2}{H_{jR} + \lambda} - \frac{G_j^2}{H_j + \lambda} \right) - \gamma \tag{15}$$

The change of the final objective function can be obtained by iteration. The formula is expressed as:

$$Obj_{split} = \frac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda} \right) - \gamma \tag{16}$$

In Equation (12), $Obj^{(s)}$ represents the total of all objective function values at leaf nodes. The difference between the objective functions of two trees before and after a node splitting is that the splitting node splits into two new nodes, while the remaining nodes remain unchanged. Therefore, the difference between the objective functions of the two trees before and after is Obj_{split} in Equation (16).

3.3. Evaluation Index Construction

For the binary classification problem, True Positive (TP), False Positive (FP), True Negative (TN), and f False Negative (FN)) are distinct combinations of the real category of data samples and the model prediction category for the binary. The two dimensions of confusion matrix are prediction class and target class. In the binary classification problem, positive and negative examples are involved. According to the actual needs of the research problem, we wish to more precisely identify enterprises with low digital level more accurately to give appropriate tilt and support in the formulation of fiscal policy. Therefore, the label of 0 (i.e., low digital level) is expressed as a positive example in this study. The confusion matrix is shown in Table 1.

Table 1. Confusion matrix.

	Positive	**Negative**
True	True Positive (TP)	True Negative (TN)
False	False Positive (FP)	False Negative (FN)

The evaluation index is defined as follows: AUC: indicates the area under the ROC curve, which is a curve with false positive rate (FPR) as the horizontal coordinate and true positive rate (TPR) as the vertical coordinate. The larger the AUC, the better the classification effect of the model. When AUC < 0.5, it indicates that the model classification effect is worse than random classification; when 0.5 < AUC < 1, the classification effect of the model is better than random classification; and when AUC = 1, the model is completely classified correctly. The indicators of the model are defined as follows.

- Accuracy is the ratio of correct prediction. The formula is expressed as:

$$\text{Accuracy } = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{17}$$

- Recall is the possibility of finding all positive samples. The formula is expressed as:

$$\text{Recall } = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{18}$$

- Precision: The correlation of classifying a negative sample as a positive is measured. The formula is expressed as:

$$\text{Precision } = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{19}$$

- F-Score considers both precision and recall indicators and is expressed as:

$$\text{F} = \frac{(\alpha^2 + 1)\text{P} * \text{R}}{\alpha^2(\text{P} + \text{R})} \tag{20}$$

- When the parameter α = 1, the most common F1 is the harmonic mean of precision and recall. The formula is expressed as:

$$\text{F1} = \frac{2{*}\text{P} * \text{R}}{\text{P} + \text{R}} \tag{21}$$

4. Results and Discussion

4.1. Feature Selection

The XGBoost algorithm's feature importance function may calculate and compare the relevance of features. Coverage, gain, and frequency are three common techniques to measure the feature importance index. Figure 3 depicts the calculation of the relevance scores of seven feature variables following the completion of data reading (The experimentally determined feature importance scores for weight, gain, and cover are identical), separation of feature variables and goal variables, and division of data sets. It reveals that the distinctive importance score for ownership is the highest. In the actual modeling process, it is vital to evaluate both the model's efficiency and its applicability. In general, the number of selected features alters the model's output. In order to increase the operational efficiency of the model when the number of features is substantial, several dimensionality reduction techniques are applied according to the actual problem. In order to verify that the future model has better generalization performance, this study traverses all features, compares the corresponding evaluation indicators, and evaluates the impact of feature count on the results.

4.2. XGBoost Model Results

In the independence test, the enterprise-related information, including enterprise region, enterprise age, ownership, number of employees, R&D investment, digital management department and national high-tech enterprise status have a significant impact on the degree of digital transformation. The XGBoost model parameters are set by the default parameters of the learner. First, all seven feature variables are included in the modeling process. From the experimental output results, it can be concluded that the AUC value of the XGBoost model that incorporates all feature variables is 0.9028, the accuracy rate is 0.9023, the recall rate is 0.9094, the accuracy rate is 0.8808, and the F1 score is 0.8948. Then, according to the ranking results of feature importance, 6, 5, 4, 3, 2, and 1 feature variables are utilized to model according to the same concept. The performance of the evaluation indicators of the XGBoost model under different feature variable selection is summarized in Table 2.

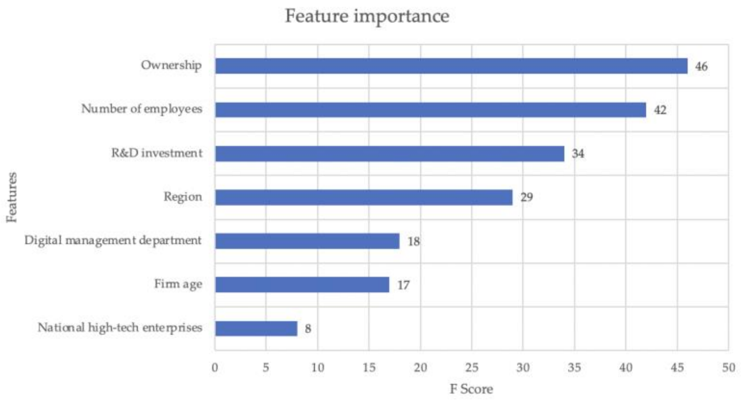


Figure 3. XGBoost feature importance score results.

Table 2. Comparison of evaluation indexes of XGBoost under different feature variable selection.

	AUC	**Accuracy**	**Recall**	**Precision**	**F1-Score**
7	0.9028	0.9023	0.9094	0.8808	0.8948
6	0.8828	0.8831	0.8796	0.8668	0.8731
5	0.8574	0.86	0.8246	0.8626	0.8432
4	0.8453	0.8451	0.8474	0.8207	0.8338
3	0.8202	0.8176	0.8529	0.774	0.8115
2	0.8153	0.8133	0.9922	0.7279	0.8299
1	0.7572	0.7558	0.7759	0.7168	0.7452

Table 2 demonstrates that the five assessment indicators of the model tend to decrease as the number of features is reduced, and when only one feature is employed, the generalization performance is poor, and the five indicators decline significantly. In addition, we can see that when all seven features are employed, the model performs best in terms of AUC, precision, and F1 score.

4.3. Grid search Parameter Tuning

Parameter tuning of the model is an integral aspect of the modeling process. The prior XGBoost model utilized the default parameters of the learner. This section focuses on parameter tuning. Common approaches for tweaking model parameters include random parameter selection, grid search parameter selection, and others. In general, good parameter tuning can improve the performance of the model. For the XGBoost model, the parameter learning rate represents the weight/learning rate of the model produced for each iteration, with a default value of 0.3. The parameter max_depth represents the depth of the decision tree, which is frequently used to control overfitting. The larger the value of the parameter gamma, the more conservative the algorithm. The parameter colsample_bytree controls the proportion of columns in each random sampling, and the parameter subsample controls the proportion of random samples taken.

After grid parameter adjustment of the XGBoost model, experimental results indicate that “colsample bytree” = 0.9, “gamma” = 0.1, “learning rate” = 0.2, “max depth” = 15, and “subsample” = 0.8 are the ideal parameter settings for the XGBoost model. The five evaluation index results following parameter tuning with all seven characteristic variables are as follows: AUC = 0.9133, Accuracy = 0.9153, Recall = 0.9254, Accuracy = 0.9052, and F1-score = 0.9185. Table 3 compares the XGBoost model before and after parameter adjustment:

Table 3. Comparison of evaluation indexes before and after parameter tuning of XGBoost model.

Evaluation Index	Before Parameter Adjustment	After Parameter Adjustment
AUC	0.9028	0.9263
Accuracy	0.9023	0.9283
Recall	0.9094	0.9384
Precision	0.8808	0.9182
F1-Score	0.8948	0.9315

4.4. Model Evaluation and Result Explanation

In the previous section, we used the XGBoost model to study the two-classification problem of enterprise digitization level and evaluated the trained classifier by constructing five evaluation indicators. In terms of comprehensive model performance, the model that retains all feature variables has superior generalization performance, but it should be noted that the model's recall rate index performs best when only the two most important features are used. Therefore, if we want to lower the level of the digital enterprises (positive example) to identify more and do not consider the accuracy (the model is the actual high level of digital enterprises identified as weak) decline, it is appropriate to feature attribute reduction and retain the two most important feature variables.

After adjusting parameters using grid search, the AUC increased from 0.9028 to 0.9263. The XGBoost model was improved to a certain extent, which shows that the selection of parameters has a great influence on the XGBoost model and that our parameter adjustment work was successful. In summary, it can be seen that the XGBoost model that is optimized by grid search parameters and retains all feature variables has better performance. Therefore, this model can be used in the digital-level binary classification study of practical problems, and the ROC curve of the XGBoost model after parameter adjustment is drawn, as shown in Figure 4.

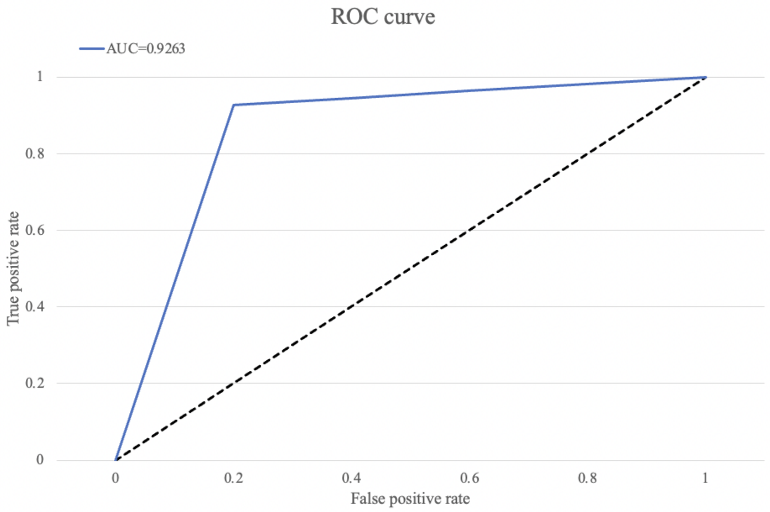


Figure 4. ROC curve of XGBoost model after parameter adjustment.

5. Conclusions

This article examines the contributing elements and degree of enterprise digitization based on the macroeconomic context of digital economy development. The XGBoost classification model is proposed. Using statistical methods and machine learning techniques, the classification of enterprise digitization level is given, along with the index system, and the corresponding index scores for the enterprise digitization levels are obtained and are an effective supplement to study the enterprise digitization level in the era of the digital economy. Specifically, we begin by extracting questionnaire data, separating the target variables from the characteristic variables, and analyzing the correlation between the target variables and the characteristic variables. The results indicate that the region, the time of establishment, the nature of ownership, the number of employees, R&D investment, national high-tech enterprise status, and the establishment of digital transformation management departments have a substantial effect on the degree of digital transformation.

In the binary classification problem of digital levels, XGBoost, an ensemble learning approach based on decision tree model, is also a representative gradient-boosting-based ensemble algorithm. We traversed the characteristic variables for each model and acquired AUC, accuracy, recall, precision, and F1-score evaluations for five classification models. The experimental findings confirm that the model containing all feature variables has superior generalization performance; hence, all features should be given priority in the actual modeling process. During the phase of model parameter adjustment, the parameters of the two models are adjusted using the grid search approach. The experimental results show that our parameter modification effort yielded certain outcomes as the model's five evaluation indices were enhanced to varying degrees, and the AUC is 0.9263.

Based on the research of this paper and the previous research results, it is evident that, on the one hand, the scale and growth process of the enterprise is an interactive process with the improvement of its digital level. Enhancing the enterprise's digital maturity can accelerate its growth. At the same time, the expansion of the scale of the enterprise also contributes to the promotion of digital transformation to a certain extent. In addition, the results of the XGBoost feature importance score indicate that the number of enterprises has a greater impact on the degree of digital transformation. On the other hand, enterprise investment in R&D funds and cooperation with scientific research institutions are crucial to promoting the integration of the digital economy and the real economy. On this basis, we believe that the following aspects should be deployed:

1. The agglomeration effect of enterprise-level digital enhancement. R&D investment and cooperation with scientific research institutes are crucial driving forces in supporting digital transformation. This notion has been validated by our empirical examination of the variable of enterprise R&D investment. The government can concentrate on establishing a digital economic industrial cluster. Enterprises in the region should aggressively participate in digital transformation. Simultaneously, efforts should be made to open up the layout of upstream and downstream industrial chains, give full play to the comparative advantages of different regions, and guide leading enterprises to unite university R&D institutions, upstream and downstream industries, and innovative consortia to form integrated industrial chain cooperation.
2. Enhance the supporting infrastructure of enterprise digital transformation. In the empirical section, it is evident that the formation of dedicated digital management departments in businesses would play a crucial role in encouraging the enhancement of the digital level. To this end, on the one hand, it is necessary to promote standardization and the standardization of the deep integration of the digital economy, to organize technical research in the field of digital economy, and to guide key enterprises to participate in the development of standards in various fields; on the other hand, it is necessary to further improve the statistical system of the digital economy, to clarify the boundary division of digital economy industry, and to strengthen the research on the compilation of industrial guidance catalogue and establish and improve the statistical classification catalogue, operation monitoring system, and other related index systems of the digital economy.
3. Accelerate the training of professionals and interdisciplinary talents in the digital economy. On the one hand, we are supposed to further improve the talent cultivation mechanism of the digital economy and increase the talent cultivation of digital technology, such as by strengthening the directional entrusted training mechanism of talents in colleges and universities; on the other hand, it is necessary to further improve the talent introduction policy, make efforts in the preferential policies for talent settlement and housing rental subsidies, and enhance the attractiveness and support accuracy of digital economy professionals.

There are still several areas for improvement in this study. The amount of data in this dataset is insufficient, to begin with. For machine learning problems, a large amount of data is generally required as support, which is more conducive to model training and learning. In the subsequent time, it is anticipated that the amount of data will increase as a result of an increase in the questionnaire's coverage and completion rate. Secondly, the category distribution of this data set is not balanced enough, which is reflected in the ineffectiveness of multi-classification research conducted with the original data. It is anticipated that the generalization performance and classification effect of the learner can still be improved by expanding the amount of data in the later stage. In addition, in terms of research methods, this study employs the XGBoost model approach for the two-classification problem of enterprise digitization level in consideration of the characteristics of the digital innovation platform, innovation combination, and innovation dissemination; research methods might also attempt to employ new research methods such as computational social science, complicated theoretical approaches, and qualitative comparative research (QCA). Finally, for the selection of features, this paper extracts seven feature variables according to the effective information in the questionnaire and conducts independence tests separately. There may be some limitations and subjectivity. In later investigations, additional features can be added and incorporated into the model.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are included within the article.

Conflicts of Interest: The authors declare that they have no competing financial or non-financial interests.

References

1. Gupta, M.; George, J.F. Toward the Development of a Big Data Analytics Capability. *Inf. Manag.* **2016**, *53*, 1049–1064. [CrossRef]
2. Vial, G. Understanding Digital Transformation: A Review and a Research Agenda. In *Managing Digital Transformation*; Taylor & Francis Group: London, UK, 2021; pp. 13–66.
3. Syverson, C. What Determines Productivity? *J. Econ. Lit.* **2011**, *49*, 326–365. [CrossRef]
4. Berman, S.L.; Wicks, A.C.; Kotha, S.; Jones, T.M. Does Stakeholder Orientation Matter? the Relationship Between Stakeholder Management Models and Firm Financial Performance. *Acad. Manag. J.* **1999**, *42*, 488–506. [CrossRef]
5. Majchrzak, A.; Markus, M.L.; Wareham, J. Designing for Digital Transformation: Lessons for Information Systems Research from the Study of ICT and Societal Challenges. *Manag. Inf. Syst. Q.* **2016**, *40*, 267–277. [CrossRef]
6. Sachs, P. Transforming Work: Collaboration, Learning, and Design. *Commun. ACM* **1995**, *38*, 36–44. [CrossRef]
7. Sebastian, I.M.; Ross, J.W.; Beath, C.; Mocker, M.; Moloney, K.G.; Fonstad, N.O. How Big Old Companies Navigate Digital Transformation. In *Strategic information management*; Routledge: London, UK, 2020; Volume 16, pp. 133–150.
8. Bharadwaj, A.; El Sawy, O.A.; Pavlou, P.A.; Venkatraman, N. Digital Business Strategy: Toward a next Generation of Insights. *Manag. Inf. Syst. Q.* **2013**, *37*, 471–482. [CrossRef]
9. Matt, C.; Hess, T.; Benlian, A. Digital Transformation Strategies. *Bus. Inf. Syst. Eng.* **2015**, *57*, 339–343. [CrossRef]
10. Gunasekaran, A.; Papadopoulos, T.; Dubey, R.; Wamba, S.F.; Childe, S.J.; Hazen, B.; Akter, S. Big Data and Predictive Analytics for Supply Chain and Organizational Performance. *J. Bus. Res.* **2017**, *70*, 308–317. [CrossRef]
11. Bellalouna, F. Digitization of Industrial Engineering Processes Using the Augmented Reality Technology: Industrial Case Studies. *Procedia CIRP* **2021**, *100*, 554–559. [CrossRef]
12. Li, W.; Liu, K.; Belitski, M.; Ghobadian, A.; O'Regan, N. E-Leadership through Strategic Alignment: An Empirical Study of Small- and Medium-Sized Enterprises in the Digital Age. *J. Inf. Technol.* **2016**, *31*, 185–206. [CrossRef]
13. Wu, Y.-Q.; Lu, H.-X.; Liao, X.-L.; Zhu, J.-M. Research on the Digitization of Manufacturing Will Enhance the Competitiveness of the Value Chain Based on Advantage Comparison. *Complexity* **2021**, *2021*, 9917772. [CrossRef]
14. Kohli, R.; Melville, N.P. Digital Innovation: A Review and Synthesis. *Bus. Inf. Syst. Eng.* **2019**, *29*, 200–223.

15. Vaia, G.; Carmel, E.; DeLone, W.; Trautsch, H.; Menichetti, F. Vehicle Telematics at an Italian Insurer: New Auto Insurance Products and a New Industry Ecosystem. *MIS Q. Exec.* **2012**, *11*, 113–125.
16. Abrell, T.; Pihlajamaa, M.; Kanto, L.; vom Brocke, J.; Uebernickel, F. The Role of Users and Customers in Digital Innovation: Insights from B2B Manufacturing Firms. *Inf. Manag.* **2016**, *53*, 324–335. [CrossRef]
17. Henfridsson, O.; Nandhakumar, J.; Scarbrough, H.; Panourgias, N. Recombination in the Open-Ended Value Landscape of Digital Innovation. *Inf. Organ.* **2018**, *28*, 89–100. [CrossRef]
18. Boersma, K.; Wagenaar, P.; Wolbers, J. Negotiating the 'Trading Zone'. Creating a Shared Information Infrastructure in the Dutch Public Safety Sector. *J. Homel. Secur. Emerg. Manag.* **2012**, *9*, 1547–7355.1965. [CrossRef]
19. Gnimpieba, Z.D.R.; Nait-Sidi-Moh, A.; Durand, D.; Fortin, J. Using Internet of Things Technologies for a Collaborative Supply Chain: Application to Tracking of Pallets and Containers. *Procedia Comput. Sci.* **2015**, *56*, 550–557. [CrossRef]
20. Warner, K.S.R.; Wäger, M. Building Dynamic Capabilities for Digital Transformation: An Ongoing Process of Strategic Renewal. *Long Range Plan.* **2019**, *52*, 326–349. [CrossRef]
21. Hazen, B.T.; Skipper, J.B.; Ezell, J.D.; Boone, C.A. Big Data and Predictive Analytics for Supply Chain Sustainability: A Theory-Driven Research Agenda. *Comput. Ind. Eng.* **2016**, *101*, 592–598. [CrossRef]
22. Wang, G.; Gunasekaran, A.; Ngai, E.W.; Papadopoulos, T. Big Data Analytics in Logistics and Supply Chain Management: Certain Investigations for Research and Applications. *Int. J. Prod. Econ.* **2016**, *176*, 98–110. [CrossRef]
23. Hess, T.; Matt, C.; Benlian, A.; Wiesböck, F. Options for Formulating a Digital Transformation Strategy. *MIS Q. Exec.* **2016**, *15*, 123–139.
24. Zhang, W.; Zhao, S.; Wan, X. Industrial Digital Transformation Strategies Based on Differential Games. *Appl. Math. Model.* **2021**, *98*, 90–108. [CrossRef]
25. Nambisan, S.; Lyytinen, K.; Majchrzak, A.; Song, M. Digital Innovation Management: Reinventing Innovation Management Research in a Digital World. *Manag. Inf. Syst. Q.* **2017**, *41*, 223–238. [CrossRef]
26. Shan, S.; Pan, J. The Effectiveness Evaluation Method of Regional Digital Economy Innovation Model Based on Intelligent Computing. *Math. Probl. Eng.* **2022**, *2022*, 8136437. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.