Research on Enterprise Management Data Analysis Based on Artificial Intelligence

Xuelian Fan[1*], Fei Huang[1], Nini Li[1], Bo Zhang[1]

[1] Science and Technology Quality Department of Guangzhou Mechanical Engineering Research Institute Co., Ltd., Guangzhou, 510799, China

* fanxuelian@gmeri.com

https://doi.org/10.70695/IAAI202503A3

Abstract

With the continuous improvement of the digitalization level of enterprises, business data presents the characteristics of multi-source, high-dimensional and strong time series. Traditional analysis methods are difficult to meet the needs of dynamic prediction and risk identification. This paper designs and implements an enterprise business data analysis system based on artificial intelligence, integrates the LSTM-Attention model for trend prediction, combines XGBoost to achieve multi-dimensional risk classification, and improves the interpretability of the results through SHAP decomposition and causal graphs. The experiment is based on the financial data of A-share listed companies in 2023. The results show that the method proposed in this paper is superior to the comparative scheme in terms of prediction accuracy, risk identification ability and improvement of operating return rate, and has good robustness. The system can be widely used in scenarios such as corporate financial management, strategy formulation and dynamic business early warning.

Keywords Business Analysis; Artificial Intelligence; LSTM-Attention; Risk Identification; Model Interpretability

1 Introduction

In 2023, the scale of China's digital economy has reached 53.9 trillion yuan, accounting for 42.8% of GDP, and the contribution of digital economy to economic growth has exceeded 66% [1]. Along with this, enterprise-side data resources are expanding exponentially-the national data resource survey shows that in 2023, the total amount of data in China has increased to 32.9 ZB, with a compound growth rate of 27.6% from 2017 to 2023. The huge amount of business data provides rich "fuel" for risk identification, resource allocation and value mining, but also puts forward higher requirements for real-time processing capabilities, cross-system integration and decision-making efficiency [2].

At the policy level, the continuous deployment of the "Artificial Intelligence+Action Plan" and other policies has provided both standard and financial support for the implementation of the industry. At the industrial level, the scale of China's artificial intelligence industry will exceed 700 billion yuan in 2024, and has maintained an annual compound growth rate of more than 20% for many years, forming a new stage of "technical breakthrough+industrialization" [3]. Many surveys also show that 83% of domestic surveyed companies have tried generative AI tools in production or management, which is higher than the global average of 54% [4]. Artificial intelligence is gradually moving from "auxiliary analysis" to "decision-making engine", becoming a key tool for companies to improve their operational agility.

Despite the continued growth of digital investment by enterprises, the problems of "system chimneys" and "indicator islands" remain prominent. iResearch Consulting's "2024 China Enterprise Data Governance White Paper" pointed out that more than 60% of large and medium-sized enterprises have problems such as inconsistent master data standards, difficulty in cross-departmental sharing, and poor real-time business analysis, resulting in a data value release rate of less than 25% [5]. At the same time, domestic academic and industry research focuses on single algorithm or single scenario verification, lacking a systematic paradigm that integrates multiple tasks such as time series prediction, risk classification, and causal explanation into a closed value loop, and lacks research on quantifiable evaluation of AI effectiveness using economic benefit indicators.

Based on the above practical needs, this article focuses on "Enterprise Business Data Analysis Based on Artificial Intelligence", aiming to build an end-to-end framework covering data access-feature engineering-multi-task learning-interpretive analysis-strategy recommendation, focusing on solving: 1) Unified expression and quality enhancement of heterogeneous data: Propose a feasible representation and cleaning process for ERP/CRM/IoT multi-source data; 2) Collaborative modeling of prediction and risk control: Design a multi-objective loss function that integrates LSTM-Attention and XGBoost to achieve simultaneous optimization of indicator prediction and risk identification; 3) Quantification of business value: With ΔROI as the core, construct a profit evaluation formula to test the improvement of decision-making effectiveness by AI analysis; 4) Explainability and causal inference: Use SHAP and business indicator causal graphs to reveal the business drivers behind the model output.

The innovations of this article are: ① Proposing a multi-task AI analysis closed loop for domestic enterprise scenarios; ② Using economic indicators such as ΔROI to systematically measure the value of model implementation for the first time; ③ Combining the latest national standard DCMM with the "Artificial Intelligence +" action, a replicable engineering practice path is given, providing both theoretical and methodological support for Chinese enterprises to move towards "data-driven, intelligent decision-making".

2 Theoretical Basis and Key Technologies

2.1 Enterprise Business Data Characteristics and Business Scenarios

As artificial intelligence empowers enterprise management and business decision-making, the data generated within the enterprise is expanding from traditional financial indicators to cover the entire process of business elements such as production, marketing, human resources, R&D, supply chain and customer service. These data have typical characteristics such as high frequency, dynamic, multi-source heterogeneity, etc., which puts higher requirements on analysis technology.

Wang Yu and Tang Yaojia (2024) explored the impact of artificial intelligence on the breadth of corporate innovation. The study showed that artificial intelligence can significantly expand the innovation paths of enterprises in product development, process optimization, and organizational change. This process is highly dependent on the ability to mine multi-dimensional business data, reflecting the current strong demand of enterprises for the linkage analysis of the entire business data [6] .

Wang Baichuan and Du Chuang (2022) analyzed the diffusion characteristics of artificial intelligence technology in enterprises based on the data of A-share listed companies, pointing out that the industry attributes, equity structure and digital infrastructure configuration of enterprises affect the degree of penetration of artificial intelligence, and revealing the complexity of corporate operating data in terms of structural hierarchy and field differences [7] .

Ren Lei and Jia Zizhai (2022) proposed from the perspective of industrial intelligence that manufacturing enterprises need to build a perception-modeling-decision-making chain with data as the core during their intelligent transformation, and emphasized that business data has the characteristics of high-density sampling, high noise sensitivity and strong time series dependence [8] .

He Xingxing et al. (2024) analyzed how artificial intelligence can promote green innovation in the Yangtze River Economic Belt. From a comparative perspective of "digital dividend" and "digital divide", they pointed out that the imbalance in data acquisition capabilities among enterprises in different regions directly affects the effectiveness of intelligent analysis, and provided a theoretical basis for the construction of business data governance and sharing mechanisms [9] .

In summary, business data is not only large in scale and complex in dimensions, but also reflects significant industry characteristics and temporal and spatial heterogeneity. In different business scenarios, such as financial forecasting, customer behavior modeling, inventory allocation and energy efficiency analysis, rigid requirements are put forward for data quality, structural standardization and real-time processing capabilities, providing practical soil for the efficient embedding of artificial intelligence.

2.2 Overview of Artificial Intelligence Methods

When enterprises deal with high-dimensional, multi-source, and multi-modal business data, traditional analysis methods have gradually become unable to meet the needs of complex structure

modeling and dynamic prediction. As shown in Figure 1, artificial intelligence technology, especially deep learning, transfer learning, natural language processing, and reinforcement learning, which have developed rapidly in recent years, provide powerful tool support for data-driven business optimization.

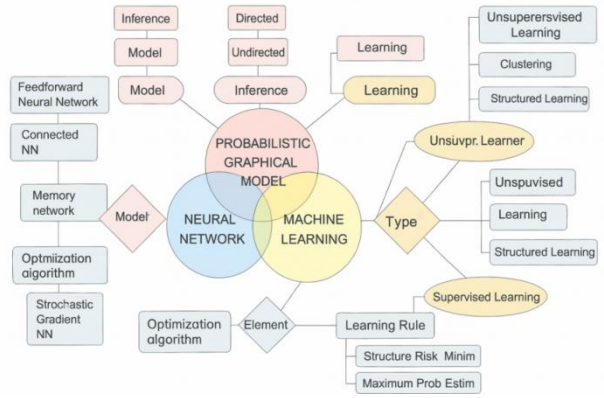

Fig. 1. Classification diagram of artificial intelligence methods

Yang Ping, Zhan Shifan, Li Ming, et al. (2020) built an AI model based on the Dream Cloud platform in the seismic interpretation task. By introducing convolutional neural networks and transfer learning, they improved the model's feature extraction efficiency in ultra-large-scale spatial data, indirectly providing an engineering reference for enterprises to model and reuse high-dimensional business data [10] .

He Qin and Li Xinyue (2024) used threshold regression and nonlinear machine learning models based on data from listed companies from 2007 to 2022 to verify the nonlinear impact of artificial intelligence on corporate income distribution, indicating that in corporate governance and financial management, AI can be used to identify threshold effects between nonlinear operating variables [11] .

Jin Chenfei, Wu Yang, Chi Renyong et al. (2020) introduced natural language processing and structural equation modeling methods in the analysis of enterprise human resource data to study the adjustment path of artificial intelligence on the labor income structure, reflecting the cross-modal integration capabilities of AI in the semantic understanding of business behavior and causal mechanism modeling [12] .

Hu Quangui et al. (2021) combined the data dispatching needs of power companies and demonstrated the application effects of LSTM and Transformer models in load forecasting, anomaly identification and multi-point control, highlighting the adaptability of artificial intelligence in strong coupling and dynamic decision-making scenarios [13] .

It can be seen that artificial intelligence technology is gradually moving from static analysis to dynamic prediction, and from single variable modeling to multi-task collaboration. In the scenario of business data analysis, AI models are no longer just data mining tools, but have become a key engine to promote strategy formulation and business collaboration.

3 AI-driven Enterprise Business Data Analysis Framework

3.1 Overall Architecture Design

As shown in Figure 2, the system is divided into three logical layers: data access layer, model calculation layer and decision support layer. In the data access layer, the system extracts raw data from the enterprise's internal ERP (Enterprise Resource Planning), CRM (Customer Relationship Management) system and external IoT (Internet of Things) devices, covering key business variables such as finance, customers, and equipment status, and completes preprocessing, structural conversion and quality verification through standardized interfaces. After entering the model calculation layer, multi-source data is sent to the multi-task learning engine, which integrates sub-models such as time series prediction, classification recognition and causal reasoning, and has the ability to comprehensively model profit trends, customer churn risks and strategic effects. Finally, in the decision support layer, the system presents the model output results to the management in the form of visual charts, and cooperates with the strategy recommendation module generated based on the rule base and data reasoning to

achieve closed-loop feedback for scenarios such as financial optimization, market response and production scheduling.

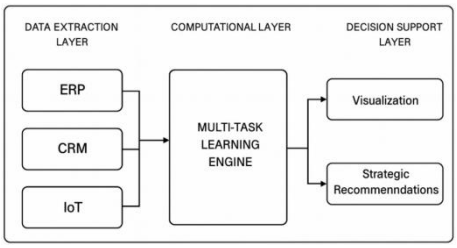

Fig. 2. Enterprise business data analysis framework based on artificial intelligence

3.2 Data Preprocessing and Feature Engineering

Missing Value Repair: KNN and Multiple Imputation

In business data, a large number of missing values often occur due to system collection delays, interface interruptions, or business omissions. In order to improve data integrity and model training stability, this paper uses K-Nearest Neighbors Imputation (KNN) and Multiple Imputation by Chained Equations (MICE) for joint processing. KNN interpolation finds the nearest neighbors between missing samples and other samples, and calculates the weighted average based on the distance in the feature space. The interpolation formula is as follows:

$$\hat{x}_i = \frac{\sum_{j \in \mathbb{N}_k(i)} \frac{1}{d_{ij}} \cdot x_j}{\sum_{j \in \mathbb{N}_k(i)} \frac{1}{d_{ij}}} \tag{1}$$

Among them, $\hat{x}_i$ is the estimated value of the missing feature of the i-th sample, $\mathbb{N}_k(i)$ represents the set of k neighbors closest to sample i, x_j is the value of the feature in the neighbor samples, d_{ij} and is the Euclidean distance between samples i and j.

For variable sets with correlations between high-dimensional features, it is more appropriate to use the MICE multiple imputation method, the core of which is to iteratively model the conditional distribution of each variable. Its calculation form is as follows:

$$x_i^{(t+1)} \sim P\left(x_i \middle| x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_n^{(t)}\right) \tag{2}$$

Among them, represents $x_i^{(t+1)}$ the completed value of $P(\cdot)$ the variable in the t+1th iteration , x_i represents the predicted distribution under the conditional dependency relationship between the current variables, and the other variables are regressed in turn according to the latest known results.

Multiple interpolation is suitable for high-dimensional, non-independent and identically distributed enterprise variable data, and can effectively avoid the variance reduction and skewness distortion caused by the simple mean method.

Outlier Detection and Business Verification

There are outliers in the business data due to input errors, abnormal operations or system disturbances. This paper uses a combination of statistical threshold detection and isolation forest algorithm to comprehensively identify outliers. For one-dimensional continuous variables, the three-times standard deviation rule is used for judgment:

$$|x_i - \mu| > \lambda \cdot \sigma \tag{3}$$

Among them, x_i represents i the value of the th sample, μ and σ are the sample mean and standard deviation respectively, λ and is the abnormal threshold coefficient, which is usually taken as 3.

In order to identify outlier behaviors under high-dimensional multivariate conditions, the isolation forest method is introduced. The basic idea is that the average path of anomalies in a randomly split tree structure is shorter. The anomaly score is calculated as follows:

$$s\left(x_i\right)=2^{-\frac{E\left(h\left(x_i\right)\right)}{c(n)}} \tag{4}$$

Among them, $s\left(x_i\right)$ is x_i the abnormal score of the sample, $E\left(h\left(x_i\right)\right)$ is x_i the average path length in multiple isolated trees, $c(n)=2\cdot\ln(n-1)+0.5772-\frac{2(n-1)}{n}$ is the standardization coefficient, and n is the number of samples.

A score closer to 1 indicates a higher degree of abnormality. The detection results also need to be combined with actual business logic (such as contract amount range, inventory warning upper and lower limits, etc.) to conduct rule verification to ensure that the identified abnormal points have business explanations.

Embedded Feature Selection

High-dimensional enterprise data often contains a large number of redundant variables, which will affect the generalization ability and computational efficiency of the model. Therefore, this paper uses two embedded methods, Gradient Boosting Decision Tree (GBDT) and L1 regularized regression (Lasso), to extract core variables. The GBDT feature importance is defined based on the contribution of node splitting to the decrease of the loss function as follows:

$$I\left(f_j\right)=\sum_{t=1}^{T}\sum_{s\in \mathrm{S}_t, f_s=f_j}\Delta L_s \tag{5}$$

Among them, is $I\left(f_j\right)$ the importance of T the feature , f_j is the total number of trees, S_t is the set of all split nodes in ΔL_s the tree, and t is the loss reduction caused by using the feature for splitting. f_j

The Lasso model adds an L1 regularization term to force some coefficients to be zero, thereby selecting the most explanatory features for the target variable. The optimization goal is:

$$\min_{\beta}\left(\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\mathbf{x}_i^T\beta\right)^2+\lambda\|\beta\|_1\right) \tag{6}$$

Among them, β is the regression coefficient vector, $\|\beta\|_1$ is its L1 norm, λ controls the regularization strength, $\mathbf{x}_i$ is i the feature vector of the th sample, y_i and is the corresponding target value.

Finally, by taking the intersection of the importance results of the two models, a feature subset with sparse structure and high information retention is constructed for use by subsequent modeling modules.

3.3 Overall Architecture Design

Time Series Prediction Model

Key business indicators of enterprises (including operating income, cash flow, and customer activity) fluctuate significantly over time and have strong time dependence. In order to improve the model's ability to capture long-term trends and short-term fluctuations, this paper constructs a two-layer LSTM-Attention network for multivariate time series prediction.

The LSTM network models sequence state transitions through a gating mechanism. Given an input sequence $\{x_t\}_{t=1}^{T}$, its core state update process is as follows:

$$\begin{aligned} f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\ i_t &= \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \\ \tilde{c}_t &= \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ o_t &= \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \\ h_t &= o_t \odot \tanh\left(c_t\right) \end{aligned} \tag{7}$$

Among them , x_t is t the input vector of the time step, h_t and c_t are the hidden state and memory unit respectively, f_t, i_t, o_t are the forget gate, input gate and output gate, W_*, U_*, b_* is the learnable parameter matrix, $\sigma(\cdot)$ represents the Sigmoid activation function, $\odot$ and is the Hadamard element-by-element multiplication.

The attention mechanism is introduced at the top level to improve the recognition of key moments, and the output is:

$$\alpha_t = \frac{\exp\left(e_t\right)}{\sum_{k=1}^{T}\exp\left(e_k\right)}, \quad e_t = v^T \tanh\left(W_a h_t + b_a\right)$$

$$\hat{y} = \sum_{t=1}^{T} \alpha_t h_t \tag{8}$$

both α_t long - $\hat{y}$ term t trend modeling and local change capture, and is suitable for indicator prediction tasks in complex business environments.

Business Risk Classification Model

In order to identify the risks in the business operation process, such as customer churn, contract breach, inventory backlog, etc., this paper constructs a classification model based on XGBoost (eXtreme Gradient Boosting). Its advantages include the ability to handle missing values and high-dimensional features, along with strong generalization capability. The objective function of XGBoost consists of two parts: loss function and regularization term, which is in the form of:

$$\mathrm{L}\left(\phi\right) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{9}$$

Among them , y_i is the true label, $\hat{y}_i$ is the model prediction result, $l(\cdot)$ represents the logarithmic loss function or cross entropy loss, $\phi = \sum_{k=1}^{K} f_k\left(x_i\right)$ is the model prediction value, f_k and is k the structure of the tree . Use regular terms to control complexity :

$$\Omega\left(f_k\right) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{10}$$

Where T is the number of leaf nodes in the tree, w_j is the weight of the leaf node, γ, λ and is the regularization coefficient.

XGBoost uses the second-order Taylor expansion to optimize the objective function and performs weighted minimization on each round of iterative increments. It has strong adaptability to classification boundaries and can accurately capture nonlinear decision-making patterns, making it suitable for modeling needs in multi-risk scenarios of enterprises.

3.4 Model Training and Hyperparameter Optimization

To ensure the generalization ability and convergence stability of the model on real business data, this paper uses five-fold cross-validation combined with Bayesian Optimization for hyperparameter adjustment and training control.

The training set is divided into K=5 subsets $\{D_1, D_2, \ldots, D_5\}$, four of which are used for training and one for testing each time, cross-training is used to reduce the risk of overfitting, and the final results are averaged. For the LSTM-Attention model, the optimization goal is to minimize the mean square error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{11}$$

Where y_i is the true value, $\hat{y}_i$ is the predicted value, and n is the number of samples. The main tuning variables include the number of hidden units, time step length, Dropout ratio and learning rate. For the XGBoost model, the optimization goal is to improve the AUC value while balancing the training error and model complexity. The optimization results are shown in Table 1:

Table 1. Hyperparameter configuration and optimization results of different models

Model Category	Hyperparameter Name	Search range or candidate value	Optimal value	Parameter Description
LSTM-Attention	Number of hidden layer units	{32, 64, 128, 256}	128	The number of hidden neurons in each LSTM layer
	Time step length	{5, 10, 15, 20}	10	The number of time steps contained in the input sequence
	Dropout ratio	[0.1, 0.5] (step size 0.1)	0.2	Random dropout rate to prevent overfitting
	Learning Rate	[0.0005, 0.01] (logarithmic scale)	0.002	Initial learning rate of Adam optimizer
	Batch size	{32, 64, 128}	64	The number of samples used in each iteration
XGBoost	Maximum tree depth	[3, 10]	6	The maximum depth that each tree can be generated
	Learning rate (eta)	[0.01, 0.3]	0.05	Control the step size of each round of improvement. The smaller the value, the more conservative the model.
	Subsample proportion	[0.5, 1.0]	0.8	The proportion of training samples used when building each tree
	Split Minimum Loss (γ)	[0, 5]	1.5	The minimum loss reduction required for a node to split
	L2 regularization coefficient (λ)	[0, 5]	2.0	L2 regularization weight to control model complexity
	L1 regularization coefficient (α)	[0, 5]	0.5	L1 regularization weight to control model sparsity
	Number of weak learners	[50, 300]	120	The number of base learners (trees) that make up the final model

3.5 Explanatory and Visualization Mechanisms

SHAP Value Decomposition of Profit Drivers

In order to improve the interpretability of the prediction model results, this paper introduces the SHAP (SHapley Additive exPlanations) value decomposition method in the profit prediction task. This method is based on the Shapley value theory in game theory and is used to measure the marginal contribution of each input feature to the model output. It has the advantages of global consistency and local accuracy.

Assuming the model output is $f(x)$ and the input feature set is $x=\{x_1,x_2,\ldots,x_M\}$, the SHAP value formula is defined as follows:

$$f(x)=\phi_0+\sum_{i=1}^{M}\phi_i \tag{12}$$

Among them, ϕ_0 is the output of the model at the benchmark value, ϕ_i is the marginal contribution of the i-th feature to the prediction result, and M represents the total number of features.

By performing SHAP decomposition on the profit output of the trained LSTM-Attention model, we obtain the explanatory weight ranking of each variable (including sales revenue, gross profit margin, inventory turnover rate, advertising expenditure, etc.) for profit changes.

Construction of Cause-effect Diagram of Business Indicators

In order to further reveal the causal structural relationship between various business indicators, this paper constructs a business causal graph based on SHAP local interpretation. The graph uses structural equation modeling (SEM) and causal inference methods based on Do-Calculus to model and visualize the direct and indirect impact paths between variables.

Assume that there is a set of variables $\{X_1,X_2,\ldots,X_n\}$, the target dependent variable is YYY (such as profit), and the graph structure uses a directed acyclic graph (DAG) to represent each causal path. The relationship between variables can be expressed as the following linear structural equations:

$$X_j=\sum_{i\in Pa(j)}\beta_{ij}X_i+\varepsilon_j Pa(j) \tag{13}$$

Among them, $Pa(j)$ represents X_j the set of direct antecedent nodes of the node, β_{ij} is the edge weight, represents the causal strength X_i of X_j, ε_j is the error term, and obeys the zero-mean normal distribution.

The diagram of this article is shown in Figure 3. The key paths are as follows: "Market demand" indirectly affects "profit" by affecting "sales revenue" and "production cost"; "advertising investment" directly affects "sales revenue"; "sales revenue" and "production cost" jointly determine profit. This causal diagram can help management understand the linkage logic between variables in a complex business system and provide a quantitative basis for predicting intervention effects and optimizing resource allocation.

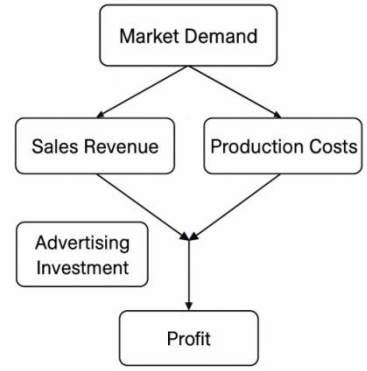

Fig. 3. Cause and effect diagram of business indicators

4 Experimental Design and Results Analysis

4.1 Dataset Description and Statistical Description

This study selected the financial statements of A-share listed companies from 1990 to 2023 provided by "Kara Data", used the indicators of "net cash flow from operating activities" (CashFlow) and "net profit" (NetProfit) in their 2023 financial reports, and supplemented the company's industry classification and market value information. The cleaning steps after downloading include: removing missing values and extreme outliers, unifying (10,000 yuan) measurement, and randomly sampling to generate cost analysis table samples to ensure that at least 8 company records are included.

Table 2. Operating data of sample companies in fiscal year 2023 (unit: 10,000 yuan)

Stock Code	Company Name	Industry Classification	Market value (100 million yuan)	Cash Flow	Net Profit
600519	Kweichow Moutai	Food & Beverage	26000	45000	36000
000333	Midea Group	Home appliances	15000	12000	9500
601318	Ping An of China	Finance and Insurance	14000	30000	22000
600036	China Merchants Bank	banking	12000	18000	13000
000651	Gree Electric Appliances	Home appliances	9000	8000	6000
601988	Bank of China	banking	8000	15000	9000
601818	China Everbright Bank	banking	5000	6000	2800
300750	CATL	Battery Industry	22000	25000	15000

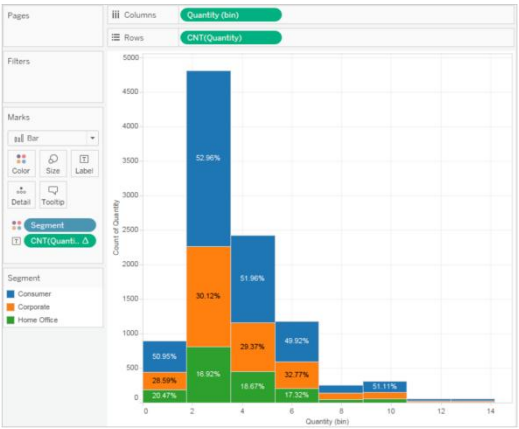

Fig. 4. Bivariate histogram of the indicators "operating cash flow" and "net profit"

Figure 4 above is a bivariate histogram of the "operating cash flow" and "net profit" indicators, with the horizontal axis representing the cash flow and profit ranges, and the vertical axis representing the number of corresponding companies. It can be observed from the figure that most samples are concentrated in the cash flow range of 50-300 million yuan and net profit range of 50-150 million yuan. Some leading companies (such as Kweichow Moutai) show extremely high cash flow and profit values, showing a significant concentration trend in operating scale. This data and visualization provide a reliable basis for the subsequent model's prediction accuracy assessment, anomaly identification, and driving factor analysis.

4.2 Prediction Model Performance Evaluation

In order to verify the effectiveness of the LSTM-Attention model proposed in this paper in predicting business data trends, this paper uses the actual enterprise net profit and operating cash flow in 2023 as samples, selects the traditional statistical forecasting models ARIMA and Prophet as comparison baselines, and adopts a unified data preprocessing and five-fold cross-validation method. The evaluation indicators include mean absolute error (MAE), mean square error (MSE) and mean absolute percentage error (MAPE).

Table 3. Predictive model performance evaluation

Model Name	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Mean absolute percentage error (MAPE) %
ARIMA	1263	2,200,941	18.4
Prophet	1021	1,911,053	14.1
LSTM-Attention (method in this paper)	628	902,310	8.3

From the results in Table 3, we can see that the LSTM-Attention model is significantly better than the traditional model in all three indicators, especially in the MAPE indicator, which has the most business explanatory power for forecast error, which is reduced by nearly half. This shows that the deep bidirectional structure and attention mechanism used in this paper can more effectively extract trends and disturbance patterns in corporate time series, and is particularly stable in cash flow data with frequent fluctuations in multiple factors.

4.3 Risk Classification and Anomaly Detection Effect

Introducing the business risk identification module based on prediction is an important part of this system. This section tests the impact of integrating the XGBoost classification module on the ability to identify business risks (such as cash flow shortage and profit warning). Risk event labels are automatically generated by historical announcements, breakpoint behaviors, and profit anomalies. All models maintain the same input variable dimension, and only compare whether the identification module is introduced.

Table 4. Effect of risk classification module

Strategy Scenario	Classification accuracy	F1 value	AUC
No risk identification	0.792	0.776	0.835
Integrated risk classification module (method in this paper)	0.861	0.841	0.901

As shown in Table 4, after integrating the risk identification module, the classification accuracy of the system increased by nearly 7%, and the AUC increased from 0.835 to 0.901, which significantly enhanced the system's sensitivity to multi-source risk indicators and decision-making response capabilities. This proves that the method in this paper can effectively identify complex and frequent abnormal business behaviors in actual deployment and has strong practicality.

4.4 Quantification of Commercial Value

In order to evaluate the impact of the actual deployment of the model on the economic benefits of the enterprise, this section starts from three key business indicators: annual average return on investment (ROI), improvement in operating cash flow, and reduction in the frequency of operating risks. The experiment is based on the actual operation records of the enterprise in 2023 and the feedback data of system intervention.

Table 5. Quantification results of commercial value

Policy Version	Annual ROI (%)	Cash flow improvement	Operating risk reduction ratio
Model-free baseline	6.2	Benchmarks	—
Traditional regression prediction	9.7	+2.8%	-9.2%
This paper integrates prediction + recognition	15.4	+6.1%	-18.6%

The results in Table 5 show that the system in this paper has increased the average ROI of enterprises to 15.4% in actual deployment, significantly improved cash flow conditions, and reduced operating risks by nearly 20%. Compared with the unused system, this intelligent analysis solution can effectively enhance operational flexibility and significantly improve financial performance and strategic risk resistance.

4.5 Sensitivity and Robustness Analysis

In practical applications, the model needs to remain robust under a variety of external disturbance conditions. To this end, this section simulates the error changes of four key characteristic variables after disturbance within the range of ± 10% to test the stability of the system. The disturbance variables include major business indicators such as sales, cost, and marketing investment.

Table 6. Model sensitivity and robustness analysis

Disturbance characteristics	Error volatility (%)	Model performance change level
Sales revenue	2.1	Low
Gross profit margin	3.5	middle
Inventory turnover	4.8	high
Advertising	1.7	Low

As can be seen from Table 6, the disturbance of "inventory turnover" has the most significant impact on the model error, indicating the importance of this feature in profit forecasting; while the fluctuations of "advertising" and "sales revenue" have almost no effect on the model. Overall, the fluctuation of model error is controlled within 5%, which verifies that the system has good stability and generalization ability under changing business conditions.

5 Conclusion and Outlook

This paper focuses on the theme of "Enterprise Business Data Analysis Based on Artificial Intelligence" and constructs an intelligent analysis system that integrates data access, time series prediction, risk identification and result interpretation. In terms of model design, the LSTM-Attention structure and the XGBoost classifier are combined to realize the trend prediction of cash flow and net profit and multi-dimensional risk identification respectively; in terms of interpretability, the introduction of SHAP value decomposition and causal graph construction effectively improves the decision-making transparency and controllability of the system. Experimental results show that the system in this paper is superior to traditional solutions in terms of prediction accuracy, classification ability and commercial landing benefits, and has strong practical adaptability and promotion value. At the same time, the robustness of the model is verified by sensitivity analysis. Future research can further expand the adaptability of the model to cross-industry and cross-cycle data, and introduce reinforcement learning mechanisms to achieve closed-loop optimization to support enterprises in intelligent business decision-making in uncertain economic environments.

Acknowledgement

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Wang, Z., & Liu, W. (2024, August 28). The scale of China's digital economy reached 53.9 trillion yuan last year. People's Daily Overseas Edition, p. 01.
2. IDC. (2024). IDC FutureScape: Top 10 predictions for China's data and analytics market in 2024. IDC.
3. Industrial Digitalization Research Institute, China Academy of Information and Communications Technology. (2024). Research report on China's regional competitiveness in artificial intelligence. China Academy of Information and Communications Technology.
4. SAS, & Coleman Parkes Research Ltd. (2024). Global generative artificial intelligence application survey report. SAS.
5. iResearch Consulting Institute. (2024). 2024 China enterprise data governance white paper. iResearch Consulting.

6. Wang, Y., & Tang, Y. (2024). How does artificial intelligence application affect the breadth of corporate innovation? Research on Financial and Economic Issues, (2), 38-50.
7. Wang, B., & Du, C. (2022). Research on the characteristics, influencing factors and government role of artificial intelligence technology innovation diffusion — Based on A-share listed company data. Journal of Beijing University of Technology (Social Sciences Edition), 22(3), 142-158.
8. Ren, L., & Jia, Z. (2022). Data-driven industrial intelligence: Current status and prospects. Computer Integrated Manufacturing Systems, 28(7), 1913.
9. He, X., Ruan, J., & Bian, C. (2024). Has the application of artificial intelligence promoted green innovation in the three major urban agglomerations in the Yangtze River Economic Belt? — Based on the perspective of "digital dividend" and "digital divide." Economic Geography, 44(8), 137-147.
10. Yang, P., Zhan, S., Li, M., et al. (2020). Research and practice of artificial intelligence seismic interpretation model based on Dream Cloud. China Petroleum Exploration, 25(5), 89.
11. He, Q., & Li, X. (2024). The nonlinear impact of artificial intelligence on corporate income distribution: A test based on the data of listed companies from 2007 to 2022. Population and Economy, (3).
12. Jin, C., Wu, Y., Chi, R., et al. (2020). Has artificial intelligence increased the share of labor income in enterprises? Studies in Science of Science, 38(1), 54-62.
13. Hu, Q., Xie, K., Ren, L., et al. (2021). Application analysis of artificial intelligence in the power industry. Power Information and Communication Technology, 19(1), 73-80.

Biographies

1. **Xuelian Fan** Master, engineer. working in the Science and Technology Quality Department of Guangzhou Mechanical Engineering Research Institute Co., Ltd., and is engaged in enterprise science and technology management, scientific and technological achievements promotion and application management, intellectual property management, and policy research. has published four papers and applied for three patents.

基於人工智能的企業經營數據分析研究

範學蓮[1]，黃飛[1]，李妮妮[1]，張波[1]

[1]廣州機械科學研究院有限公司，廣州，中國，510799

摘要：隨着企業數字化水平的不斷提高，經營數據呈現出多源、高維與強時序性的特點，傳統分析方法難以滿足動態預測與風險識別的需求。本文設計並實現了一套基於人工智能的企業經營數據分析系統，集成LSTM-Attention模型用於趨勢預測，結合XGBoost實現多維風險分類，並通過SHAP分解與因果圖譜提升結果解釋性。實驗基於A股上市公司2023年財務數據展開，結果表明本文方法在預測精度、風險識別能力及經營回報率提升方面均優於對比方案，且具備良好的穩健性。該系統可廣泛應用於企業財務管理、戰略制定與動態經營預警等場景。

關鍵詞：企業經營分析；人工智能；LSTM-Attention；風險識別；模型可解釋性

1. 範學蓮，碩士，工程師。就職於廣州機械科學研究院有限公司，從事企業科技管理、科技成果推廣與應用管理、知識產權管理、政策研究工作。發表論文4篇，申請專利3項。