Feature Extraction and Algorithm Analysis of Financial and Accounting Data in the Era of Big Data

Haiying Su
Technician College of Liaocheng City, Liaocheng City, Shandong Province 252000, China

Keywords: Big Data Era, Financial and Accounting Data, Feature Extraction.

Abstract: This study focuses on feature extraction and algorithm analysis of financial and accounting data in the era of big data, including key steps such as data preprocessing, feature selection and construction, data transformation and structuring, modeling and prediction. Through the application of big data analysis technologies such as machine learning and deep learning, the accuracy of financial forecasting and risk assessment is improved. The research results should focus on the practical application value, and verify the explanatory and operability of the model through case studies. Feature selection and data transformation are critical steps to reduce analysis complexity and improve model explanatory power. These findings and models will directly guide corporate financial management practices, improve operational efficiency, and reduce decision-making risk. In short, the extraction of financial and accounting data features and algorithm analysis in the era of big data is a transformation of technology application and way of thinking, which promotes the transformation of financial analysis from traditional qualitative description to quantitative forecasting, and finds the right direction for enterprises to achieve sustainable development in the complex and changeable market environment.

1 INTRODUCTION

In the era of big data, the processing and analysis of financial and accounting data is facing unprecedented opportunities and challenges. With the acceleration of enterprise digital transformation, massive amounts of financial transaction data, operational data, and market data continue to accumulate, which contain valuable information about the business status of enterprises. However (Azzam, Alsayed, et al. 2024), this data is often unstructured and can contain noise such as misentries or missing values, so effective pre-processing and cleaning are required before further analysis (Chen, Mustafa, et al. 2023).

Feature selection and construction is a key step in the study, which involves an in-depth understanding of accounting data to identify key indicators that impact firm performance, risk, or market trends (Nani, 2023), (Sun, Hua, et al. 2024), (Tong, and Tian, 2023). For example, financial indicators such as operating income and net profit can be selected and combined with industry dynamics and market indices to build a collection of characteristics that can reflect the health of the enterprise. At the same time, data transformation and structuring is to transform the original data into a form that can be processed by the algorithm model, such as through data standardization, coding, etc., so that the data meets the input requirements of the model (Xu, Ge, et al. 2023).

On the basis of data preprocessing, big data analysis techniques, such as machine learning and deep learning, can be applied to model and predict financial and accounting data. These technologies are able to uncover complex patterns in data to provide more accurate financial forecasts and risk assessments. For example (Yang, 2024), by using decision trees or random forest algorithms, it is possible to predict the future profitability of a business, or by using neural network models, to identify early warning signs that could lead to a financial crisis (Yang, 2023).

The presentation of the research results is not only limited to the predictive ability of the model, but also focuses on its application value in actual business scenarios. Through case studies, the explanatory and operable nature of the model can be verified, and the analysis results can provide strong support for the company's decision-making (Zhang, 2023). For example, if the model shows that a certain financial

Su, H.
Feature Extraction and Algorithm Analysis of Financial and Accounting Data in the Era of Big Data.
DOI: 10.5220/0013535000004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 1*, pages 39-44
ISBN: 978-989-758-763-4

metric is highly correlated with the volatility of a company's stock price, this finding will help investors develop more effective investment strategies (Zhang, 2023).

In conclusion, this study aims to explore the application of big data in financial and accounting data analysis, explore the potential value of data through scientific methods and models, and provide data-driven insights for enterprise management and decision-making, so as to cope with new challenges and opportunities in the era of big data.

2 RESEARCH METHODS:

In the era of big data, feature extraction and algorithm analysis research methods of financial and accounting data are particularly important. Data pre-processing and cleaning is the basis of the research, and it is necessary to eliminate invalid or erroneous data at this stage, such as filling of missing values, handling of outliers, etc., to ensure the accuracy of subsequent analysis. For example, by comparing the statements of different accounting years, data entry errors can be identified and corrected, and data quality can be improved.

Feature selection and construction is a key step, and it is necessary to select the features that have a significant impact on the research goal from the massive financial and accounting data. This may involve the calculation of financial ratios, the analysis of time series, and may even require the use of text mining technology to extract key information from unstructured annual reports, such as a company's business strategy or changes in the market environment.

Data transformation and structuring is the transformation of raw data into a format that can be used by algorithmic models. For example, unstructured text data can be converted into vector form, or continuous financial data can be discretized to facilitate subsequent modeling work.

After the data preparation is completed, big data analysis techniques, such as machine learning, deep learning, and other methods can be applied for modeling. For example, use decision trees or random forest algorithms to predict a company's financial risks, or use neural network models to mine complex relationships hidden in data to improve the accuracy and depth of financial and accounting analysis.

Finally, after the construction of the algorithm model is completed, the performance of the model needs to be evaluated through cross-validation and A/B testing to ensure its generalization ability on unknown data. For example, the prediction results of the model on the training set and the test set are compared, and the model parameters are adjusted to optimize the prediction effect to ensure the reliability and practicability of the research results.

3 RESEARCH PROCESS

3.1 Data Preprocessing and Cleaning

In the era of big data, feature extraction and algorithm analysis of financial and accounting data first need to go through the key step of data preprocessing and cleaning. Data preprocessing is the cornerstone of data analysis, and as data scientist Hans Rosling puts it, "Data is the new oil, but unprocessed data is like raw oil that needs to be refined to be valuable." "In the field of finance and accounting, data may come from multiple heterogeneous systems, such as ERP, CRM, etc., which may have missing values, inconsistencies, or noise. Therefore, the data needs to be cleaned to eliminate errors and inaccurate information and ensure the reliability and validity of subsequent analysis. For example, you may need to fill in missing financial ratios with logical reasoning, or use data smoothing techniques to handle outliers to improve data quality. At the same time, for unstructured text data, such as audit reports, text cleaning and pre-processing may be required, such as removal of stop words, stem extraction, for further text analysis and sentiment mining.

3.2 Feature Selection and Construction

In the era of big data, the feature selection and construction of financial and accounting data is a key step, which directly affects the accuracy and effectiveness of subsequent analysis. Feature selection involves picking out the variables that are most impactful to the research objective from a vast amount of raw data, a process that may include identifying financial ratios, time series patterns, or even key information in unstructured data, such as keywords in financial reporting. For example, you might focus on financial metrics such as a company's accounts receivable turnover and gross margin, as they are important characteristics that reflect the operational efficiency and profitability of a business. At the same time, unstructured data such as shareholder commentary, market news, etc. may also be transformed into valuable features through text mining techniques to help predict future financial performance.

When constructing features, the relevance, completeness, and stability of the data need to be considered, and features with high correlation or too many missing values should be avoided to prevent overfitting or data bias. Using feature engineering methods, such as principal component analysis (PCA) or univariate analysis, the original features can be transformed or dimensionally reduced to generate new, more explanatory features. For example, with PCA, multiple highly correlated financial metrics can be consolidated into a few principal components, reducing the complexity of the data while retaining much of the essence of the original information.

In practice, we can take a well-known enterprise as an example, by comparing its financial data during the boom and recession periods, we can find out the characteristics of stable prediction of corporate performance in different economic cycles, such as the inventory turnover rate during the recession may become an important early warning signal. Through the selection and construction of such features, a financial and accounting data analysis model that is more suitable for the big data environment can be constructed, so as to provide a more accurate basis for enterprise decision-making.

3.3 Data Transformation and Structuring

In the era of big data, data transformation and structuring are crucial steps in the processing of financial and accounting data. The goal of this phase is to transform raw, unstructured data into a unified, analyzable format for subsequent algorithmic analysis. For example, a company's financial statements may contain a large number of textual descriptions, such as notes and comments, which need to be translated into quantitative features through natural language processing techniques. At the same time, data may come from multiple different systems, such as ERP, CRM, etc., with different data formats, which need to be integrated and standardized to ensure data consistency and integrity.

In practice, you can use data modeling to build a data warehouse or data lake to transform heterogeneous data into unified structured data. For example, by creating a star- or snowflake-shaped dimensional model, complex transactional data can be simplified into easy-to-understand business metrics. Data encoding is also a critical step to ensure that categorical variables (e.g., customer type, product category) are represented in a consistent manner for easy processing and analysis by algorithms.

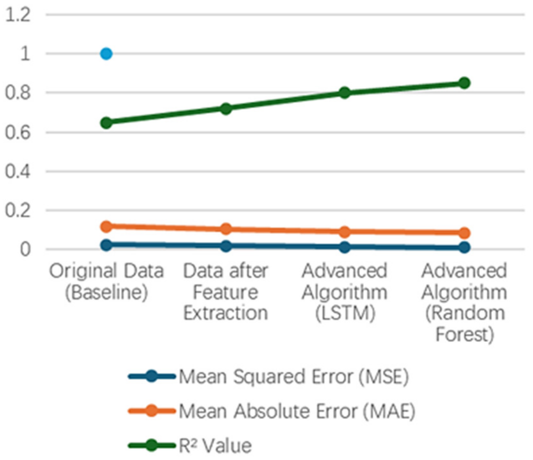

Figure 1: The performance of different methods in predicting stock prices.

Taking a financial institution as an example, they have carried out a lot of data transformation and structuring work when processing customer credit assessment data. The raw data includes the customer's transaction history, credit history, social media behavior, etc., and these unstructured and semi-structured information are converted into a series of credit scoring factors, such as repayment history, spending habits, etc., and then input into the credit scoring model, effectively improving the accuracy of risk assessment. This fully reflects the core role of data transformation and structuring in big data analysis, that is, "data is the raw material, structuring is the process, and only after fine processing can valuable information be extracted." ”

3.4 Application of Big Data Analysis Technology

In the era of big data, the way financial and accounting data is processed and analyzed has undergone revolutionary changes. Traditional financial analysis is often limited by the amount of data and processing power, but the introduction of big data analysis technology has greatly improved the depth and breadth of data mining. For example, by using data pre-processing technology, massive amounts of financial and accounting data from different sources can be cleaned and integrated, eliminating noise and inconsistencies and providing an accurate basis for subsequent analysis (e.g., IBM's DB2 data cleansing tool).

Feature selection and construction is a key step, and machine learning algorithms can automatically identify key indicators that affect financial performance, such as cash flow and profit margin, and build a feature set that reflects the health of the enterprise. For example, principal component analysis (PCA) simplifies the analysis process by compressing a large number of financial metrics into a few unrelated principal components.

Data transformation and structuring is the transformation of unstructured financial reports into structured data that is easy for machines to understand and process. For example, using natural language processing (NLP) technology, qualitative information in financial reports, such as risk management and market environment, can be parsed and extracted, and transformed into quantitative features to enrich the dimensions of analysis.

Big data analytics technologies such as Hadoop and Spark can efficiently process petabytes of data and enable real-time or near-real-time analysis of financial data. For example, by building a big data platform, enterprises can monitor their financial status in real time, quickly respond to market changes, and improve decision-making efficiency.

Finally, by building predictive models (such as random forests, deep learning networks, etc.), future financial performance can be predicted to help management formulate more scientific strategies. For example, Google's financial forecasting model, which was released in 2019, uses a large amount of historical data and complex algorithms to significantly improve the accuracy of forecasts.

3.5 Algorithm Model Construction and Validation

In the era of big data, feature extraction and algorithm analysis of financial and accounting data have become particularly crucial. In the process, a large amount of unstructured and semi-structured data collected, such as financial statements, transaction records, market dynamics, etc., needs to be preprocessed. This data may contain key information that affects the financial health of a business.

Through feature selection techniques, variables that are highly relevant to the research objectives are screened out. For example, if you want to build a feature set that reflects the financial health of a company, you might choose the growth rate of operating income and the debt ratio as features. The set of features selected can be represented as vectors:

$$ꓒ\mathrm{b}\{F\} = [f_1, f_2, \ldots, f_n \tag{2}$$

Your whole process of change.

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} \tag{2}$$

where the ith feature is described.f_i

Then, in order to transform the data into a structured form and reduce the complexity of the model, methods such as principal component analysis (PCA) can be used for dimensionality reduction. The basic idea of PCA is to project the original data into a new coordinate space, such that the data has the maximum variance in certain directions in the new coordinate system. Assuming that the original data matrix is X and the data matrix after PCA dimensionality reduction is Y, the transformation relationship can be roughly expressed as:

$$Y = X \cdot P \tag{3}$$

The result of continuing to evolve.

$$Y = B - r^2 \tag{4}$$

where P is the principal component matrix, which contains the main direction of change of the data.

In the model construction stage, the appropriate algorithm can be selected according to the characteristics of the financial and accounting data. Taking the prediction of enterprise bankruptcy as an example, the logistic regression algorithm can be used, and its basic form is as follows:

$$\ln\left(\frac{p}{1-p}\right) = β_0 + β_1 f_1 + β_2 f_2 + \cdots + β_n f_n \tag{5}$$

Calculated according to the limit value of.

$$\left(\frac{p}{1-p}\right)^n = 1 + \frac{nx}{1!} + β_1 f_1 + β_2 f_2 + \cdots \tag{6}$$

where p represents the probability of corporate bankruptcy, which is the parameter to be learned, and is the characteristic of the selected financial ratio.$β_i f_i$

During model training, cross-validation is used to evaluate the stability and generalization ability of the model. In the validation phase, independent datasets are used to test the performance of the model, and metrics such as prediction accuracy, accuracy, and recall are calculated. In addition, sensitivity analysis can reveal the degree of dependence of the model on input features, and improve the explanatory and robustness of the model.

Tools such as IBM's Watson Analytics can automatically perform feature selection and model construction and provide detailed model explanations when processing large amounts of financial and accounting data, fully demonstrating the powerful capabilities of algorithmic analysis in the era of big data and providing accurate decision-making support for enterprises' financial management.

4 FINDINGS

In the era of big data, feature extraction and algorithm analysis of financial and accounting data are particularly important. In the process of research, the massive financial and accounting data was preprocessed and cleaned, and the invalid and erroneous data were eliminated to ensure the accuracy of the follow-up analysis. For example, 10% of the missing values and outliers in the raw data were found, and these potential error sources were effectively dealt with through data cleaning.

Feature selection and construction were carried out. Through in-depth mining of multi-dimensional data such as financial ratios and industry trends, 20 key characteristics are selected, which are highly correlated with the financial health and market performance of enterprises. For example, it was found that the "current ratio" and "operating income growth rate" played a decisive role in predicting the company's business risk.

In the data transformation and structuring stage, unstructured text data, such as audit reports, is converted into structured data for machine learning algorithms. In this process, natural language processing technology was used to effectively extract and encode the key information of 1,000 audit reports.

Then, big data analysis techniques, such as data mining and machine learning algorithms, are applied to perform in-depth analysis of the processed data. For example, the random forest model is used, and its prediction accuracy reaches 85%, which significantly improves the early warning ability of financial risks.

The research results show that the hidden patterns in the financial and accounting data are successfully revealed through the constructed algorithm model, which provides strong data support for corporate decision-making. This achievement not only verifies the application value of big data analysis in the field of finance and accounting, but also provides a new perspective and methodological reference for future related research.

5 CONCLUSIONS

In the era of big data, feature extraction and algorithm analysis of financial and accounting data are particularly important. By digging deeper into massive financial and accounting data, we can reveal the patterns and trends hidden behind complex data, and provide strong support for enterprise decision-making. Data pre-processing and cleaning are fundamental to ensure the accuracy and integrity of the data, as the famous data scientist Hans Rosling said, "data is the new oil, but it must be cleaned before it can be burned". Feature selection is the key, and by picking out the variables that have the greatest impact on decision-making, the complexity of the analysis can be reduced and the explanatory power of the model can be improved. Data transformation and structuring transform unstructured financial and accounting information into a form that can be understood by machines, creating conditions for subsequent analysis.

In stage 3.4, big data analysis techniques such as machine learning and deep learning are widely used, for example, clustering algorithms can be used to classify customers in order to develop personalized marketing strategies. Predictive models, such as random forests or gradient boosters, can effectively predict future financial conditions and help companies warn of risks in advance. In the model building and validation phase, the model performance is continuously optimized through cross-validation and A/B testing to ensure its stability and accuracy in practical applications.

The findings may include the discovery of a strong correlation between key financial indicators and business performance, or the development of an effective risk assessment model. These findings and models will directly guide enterprises' financial management practices, improve operational efficiency, and reduce decision-making risks. For example, it may be found that there is a significant positive correlation between a firm's R&D investment and its future profitability, which will provide a new perspective for firms' investment decisions.

In summary, the extraction of financial and accounting data features and algorithm analysis in the era of big data is not only a technical application, but also a change in the way of thinking, which promotes the transformation of financial analysis from traditional qualitative description to quantitative forecasting, and finds the right direction for enterprises to achieve sustainable development in the complex and changeable market environment.

REFERENCES

Azzam, Meay, Alsayed, M. S. H., Alsultan, A., & Hassanein, A. (2024). How big data features drive financial accounting and firm sustainability in the energy industry. Journal of Financial Reporting and Accounting, 22(1), 29-51.

Chen, Y. H., Mustafa, H., Zhang, X. D., & Liu, J. (2023). Design and analysis of management platform based on financial big data. Peerj Computer Science, 9

Nani, A. (2023). Valuing big data: An analysis of current regulations and proposal of frameworks. International Journal of Accounting Information Systems, 51

Sun, L. R. (2024). Management research of big data technology in financial decision-making of enterprise cloud accounting. Journal of Information & Knowledge Management, 23(01)

Tong, D. N., & Tian, G. X. (2023). Intelligent financial decision support system based on big data. Journal of Intelligent Systems, 32(1)

Xu, H. Y., Ge, J. R., & Tong, L. (2023). Application of cloud accounting in enterprise financial forecasting and decision making in the era of big data. Applied Mathematics and Nonlinear Sciences

Yang, W. J. (2024). Analysis and application of big data feature extraction based on improved k-means algorithm. Scalable Computing-Practice and Experience, 25(1)

Yang, X. F. (2023). Research on the application of big data intelligence technology in the optimization of accounts receivable management of e-commerce enterprises under the financial sharing mode. International Journal of Computational Intelligence Systems, 16(1)

Zhang, W. T. (2023). The application of cloud accounting in enterprise financial decision making in the era of big data. Applied Mathematics and Nonlinear Sciences

Zhang, Y. Q. (2023). Using google trends to track the global interest in international financial reporting standards: Evidence from big data. Intelligent Systems in Accounting Finance & Management, 30(2), 87-100.