International Journal of Computer Applications Technology and Research
Volume 12–Issue 11, 05 - 11, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1211.1002

Interpolation Fusion Strategy with LSTM for Tax Data Prediction: An Application

Daifang Gou
College of Communication Engineering,
Chengdu University of Information Technology,
State Administration of Taxation Yibin Sanjiang New District Taxation Bureau,
Chengdu, China

Chengyu Hou
College of Communication Engineering,
Chengdu University of Information Technology,
Network and Data Security Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China,
Chengdu, China

Linsong Xiao
College of Communication Engineering,
Chengdu University of Information Technology,
Network and Data Security Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China,
Chengdu, China

Chang Liu
Chengdu Civil Affairs Bureau Office,
Chengdu, China

Hantao Liu
Education Informationization and Big Data Center, Education Department of Sichuan Province,
Chengdu, China

Wenzao Li
College of Communication Engineering,
Chengdu University of Information Technology,
Network and Data Security Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China,
Chengdu, China

Abstract: Traditional data analysis methods have limitations that result in many valuable tax insights being overlooked in routine tax data processing. However, with machine learning techniques, it is feasible to uncover insights that traditional methods fail to capture. Establishing a correlation between historical and future tax data has been a challenging topic, with no effective methods for resolution. Although various models, based on linear and non-linear data, exist for forecasting, they often produce significant errors when applied to tax data predictions.

This paper addresses the non-linear characteristics and volatility of tax data, proposing an enhanced Long Short-Term Memory (LSTM) model. In contrast to the conventional LSTM model, this improved model boasts a higher prediction accuracy. The enhancement involves interpolating the input data and fusing the interpolated data back into the original dataset, aiming to augment the accuracy of the output. In our experimental phase, genuine tax data was used for forecasting, and the superiority of the enhanced LSTM model over the traditional one was visually demonstrated through charts.

Upon predicting tax data for two companies and comparing the outcomes to actual scenarios, it was observed that the proposed enhanced LSTM model significantly reduced the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) by 13.65% and 14.49%, respectively, compared to the traditional LSTM model. This indicates the distinct advantage of the improved LSTM model in enhancing the accuracy of tax data predictions.

Keywords: data analysis methods, tax data, LSTM, accuracy,data predictions.

1. INTRODUCTION

1.1 Research Background

Utilizing the LSTM model to analyze tax data introduces an approach harnessing the power of Long Short-Term Memory neural networks to process and interpret data across diverse scenarios [1]. In today's society, the handling and analysis of tax data have become increasingly critical [2]. Tax data, brimming with valuable insights, holds significance for governmental departments, enterprises, and research institutions. However, traditional methods of analyzing tax data present certain limitations, preventing many important tax insights from being fully leveraged. Concurrently, the rise of contemporary technologies like machine learning and deep learning offers new methodologies for a better understanding and analysis of tax data.

Compared to traditional data analysis methods, the LSTM model showcases notable advantages in predicting tax data. Representing a hallmark of deep learning, the LSTM model, equipped with superior sequential modeling capabilities, effectively captures temporal dependencies in tax data, aiding precise future tax trend and variation predictions [3]. Being a nonlinear model, the LSTM contrasts traditional linear models by aptly handling complex nonlinear relations in tax data, thereby enhancing prediction accuracy and reliability [4]. Given the massive and diverse nature of tax data, the LSTM model, with its parallel computational abilities and versatility across data types, efficiently manages large datasets, facilitating comprehensive tax data analysis. Furthermore, its capability to seize and utilize long-term data dependencies is pivotal for analyzing prolonged trends and cyclical shifts in tax data. The real-time and adaptive features of the LSTM model enable on-the-fly parameter updates to navigate the evolving tax environment and influx of new data, proving to be highly adaptable in practical applications.

1.2 Existing Challenges

Traditional tax data analysis techniques are riddled with significant limitations, encompassing data complexity, handling nonlinear relations, capturing long-term trends, and real-time responsiveness. Tax data, inherently complex, consists of a plethora of data types and features. Conventional methods often falter in effectively addressing this data diversity, resulting in information loss or diminished analysis accuracy. Tax data embodies intricate nonlinear relations, potentially influenced by multifaceted factors like economic fluctuations and policy shifts. Classical linear models fall short in grasping these sophisticated nonlinear dynamics, leading to errors in predictions and analyses. Additionally, capturing long-term trends and cyclical changes, essential in tax data, remains a challenge for traditional techniques. Real-time analysis stands as one of the linchpins in tax data interpretation; with a continuous stream of new data, the analytical approach necessitates real-time adaptability to maintain accuracy and relevance [5].

1.3 Proposed Solution

To address the aforementioned challenges, this paper presents an enhanced LSTM network model tailored for tax data analysis. Considering the nonlinear attributes and volatility of tax data, this refined model, compared to its traditional LSTM counterpart, boasts heightened prediction accuracy. The enhancement involves interpolating the model's input data and integrating the post-interpolation data back into the primary dataset, aiming to elevate output data accuracy. Through this methodology, the complex nonlinear relationships within tax data can be better understood, facilitating more precise future tax trend predictions. With these advancements, tax data can be analyzed more effectively, yielding accurate forecasts and profound insights.

1.4 Paper Structure

This paper proposes an improved LSTM model, demonstrating superior accuracy in real-world tax data prediction compared to the conventional LSTM model. The paper's contributions are as follows:

(1) We employ the traditional LSTM model to forecast based on real tax data, presenting the respective predictive results visually in Chapter 4. Comparative results of predictions against actuals, as well as evaluation metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), are discussed [6,7].

(2) Through interpolation of input data and integrating this interpolated dataset as a segment of the primary input, an advanced LSTM model is derived and applied for predictions. Final comparative results, along with evaluation metrics MAE and RMSE, are outlined, juxtaposing the accuracy with that of the conventional LSTM model.

(3) Conclusively, performance evaluations of both LSTM models are conducted. Chapter 6 culminates with an in-depth summary of the characteristics of both LSTM models, emphasizing the merits of the proposed enhanced LSTM model and its quantified advantages.

2. RELATED WORK

In this chapter, we mainly introduce traditional data analysis and forecasting methods and delve into the application of machine learning in tax data prediction. We highlight the current challenges and explore the latest methods to address these challenges.

2.1 Traditional Data Analysis Methods

In the realm of traditional data analysis, the processing of tax data often relies on conventional methods from fields such as statistics and economics. These methods encompass time series analysis, regression analysis, and hypothesis testing [8,9,10]. Time series analysis is extensively utilized to discern temporal trends and cyclical variations in tax data, yet it generally struggles with handling intricate nonlinear relationships. Regression analysis is frequently employed to investigate causal relationships in tax data, but its capability to model interactions among multiple variables and nonlinear relationships is limited. Hypothesis testing methods are used to verify assumptions about tax data, but they may overlook the intricacies and uncertainties inherent in the data.

2.2 Application of Machine Learning in Tax Data Prediction

With the advent of machine learning and deep learning, an increasing number of studies have started to explore their applications in tax data analysis. Machine learning models, such as decision trees, random forests, and support vector machines, have been utilized for the classification and prediction of tax data [11,12,13]. They serve purposes like classification, regression, and forecasting in the context of tax data. Compared to traditional methods, machine learning models are adept at capturing complex relationships within the data, thereby enhancing the accuracy and reliability of tax predictions. However, they still grapple with dependencies on data feature engineering and challenges in handling time series data.

3. LSTM MODEL AND MODELING BASED ON TAX DATA

In the previous chapter, we discussed traditional data analysis methods and the application of machine learning in tax data prediction. In this chapter, we delve deeper into the LSTM model and its modeling based on tax data, including both traditional LSTM models and modified versions. We will also introduce methods of numerical interpolation processing and discuss how to train and forecast using these models to better comprehend how to utilize them to enhance the prediction accuracy and reliability of tax data.

3.1 LSTM Model Based on Tax Data

In the context of tax data, time series data can be visualized as sequences with continuous time steps, where each step encompasses an observation, such as tax revenue or economic indicators. In this section, we explore the traditional LSTM model tailored for tax data, which is a variant of Recurrent Neural Networks (RNN) and is widely used for sequence data modeling [14]. The LSTM model aims to address challenges faced by conventional RNNs when dealing with long sequence data and long-term dependencies. To counteract the shortcomings of RNNs, LSTM introduces a specialized gating mechanism, comprising input gates, forget gates, and output gates, along with internal memory cells [15]. The role of these gates is to control the flow of information, allowing LSTM to capture long-term dependencies more effectively. The forget gate permits the model to selectively forget previous information, while the input and output gates decide when to introduce new information and when to produce outputs. These mechanisms help alleviate the vanishing gradient problem while retaining sensitivity to long-term dependencies.

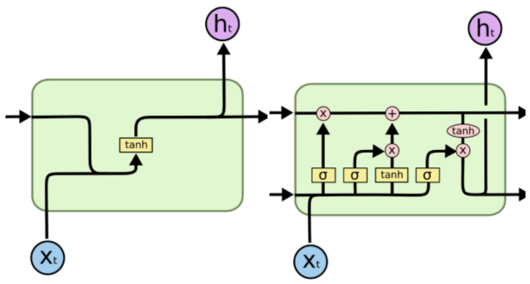


Figure. 1Comparison between LSTM model and traditional RNN model

The architecture of a traditional LSTM (Long Short-Term Memory) model includes an input layer, hidden layers, and an output layer. The input layer typically consists of input nodes representing time steps and other relevant features. The hidden layers are composed of multiple LSTM units, each of which has its own internal state (memory cell) and output gate. The design of these LSTM units allows them to capture long-term dependencies in sequential data, which helps in making more accurate predictions of future trends and changes.

The training process of a traditional LSTM model typically involves steps such as data preparation, model construction, model training, and model prediction. First, time series data is prepared in a format suitable for model input. Then, an LSTM model is constructed with a clear definition of its architecture and parameters. Subsequently, the model is trained using historical tax data by fine-tuning its weights and parameters through the minimization of a loss function to better fit the data. Finally, the trained LSTM model can be used for predicting future tax data. Although traditional LSTM models perform well in handling time series data, they may encounter performance bottlenecks when dealing with complex nonlinear relationships and noisy data. Therefore, the development of improved LSTM models to enhance performance becomes necessary.

3.2 Improved LSTM Model Based on Tax Data

Building upon the traditional LSTM model, we have introduced an enhancement known as data interpolation, which demonstrates improved performance when dealing with tax data [16]. In this enhanced LSTM model, we calculate the average of the past two time steps' data and insert this average as a new data point into the input data. This process is referred to as data interpolation, and its purpose is to smooth the original data, reduce noise in the data, and enhance the training and prediction capabilities of the model. Data interpolation aids in improving the LSTM model's ability to capture the characteristics of the data when dealing with complex nonlinear relationships. By introducing the average value, we can reduce spikes and fluctuations in the data, enabling the model to learn the data patterns more stably. This is particularly useful when handling time series data like tax data, which exhibits periodicity and trends. The improved LSTM model can more accurately capture these trends and periodic changes, thereby enhancing the accuracy of tax data predictions.

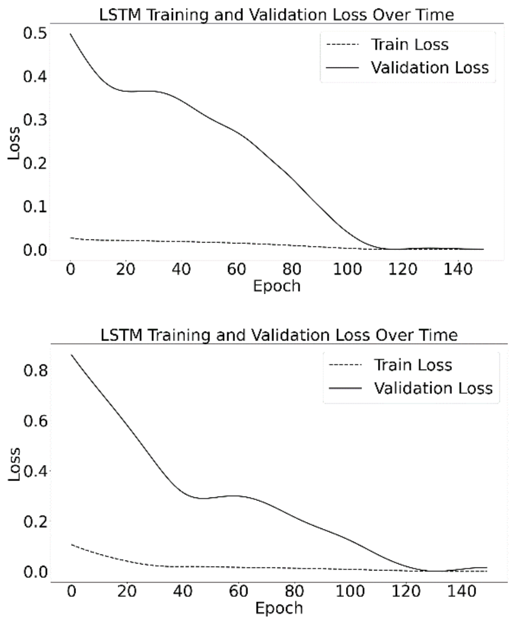


Figure. 2Loss Comparison between LSTM Model and Improved LSTM Model

Figure 2 displays the loss curves during the training process of the traditional LSTM model and the improved LSTM model. It can be observed that the loss trends of both models are similar, but the improved LSTM model exhibits greater stability in some cases. This indicates that the improved model fits the data better during training, reducing the influence of fluctuations and noise. By introducing the enhancement of data interpolation, we can make better use of the information in tax data, gain a deeper understanding of the nonlinear relationships within the data, and enhance the model's performance.

3.3 Model Training and Prediction

In the process of building the improved LSTM model, the training and prediction steps are of critical importance, forming the foundation for the final tax data predictions. Model training is a complex process that begins with data preparation, including cleaning and preprocessing the raw tax time series data to ensure its suitability for input into the LSTM model. The model construction phase involves defining the LSTM architecture and selecting its parameters, including input dimensions, the number of hidden units, and output dimensions. The choice of the loss function and optimizer is also crucial; the appropriate loss function directly affects the model's performance, while the optimizer selection impacts the speed and stability of the training process.

For the input dimension, we selected 1 to represent a single feature (taxable amount), and the hidden dimension (i.e., the number of hidden units) was set to 50. We used Mean Squared Error (MSE) as the loss function and the Adaptive Moment Estimation (Adam) optimizer to optimize the model's weights and parameters. Using the trained LSTM model, we performed sliding window predictions on the entire dataset to obtain future predictions.

Regarding the prediction results, we conducted experimental validation using real tax data from XX city for the years 2019 to 2020. We used the taxable amount for the corresponding months in these two years to predict the taxable amount for each month in 2021. We performed predictions using both the LSTM model and the improved LSTM model, and we visualized the differences between the two models using line graphs and bar charts for Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The improvement in performance was quantified by the difference in MAE and RMSE values.

4. SIMULATION AND DISCUSSION

In this chapter, we will provide a detailed overview of the experimental setup, present the results, and discuss the implications of the experimental findings. Firstly, we will outline the experimental settings, including data sources and the specific objectives of the prediction task. Subsequently, we will showcase the experimental results and quantitatively evaluate the performance improvement of the enhanced LSTM model using metrics such as MAE and RMSE.

4.1 Experimental Setup

The experiments in this study are based on real tax data sourced from XX city's tax records. The objective of the experiments is to use historical tax data (data from corresponding months in 2019 to 2020) to predict future tax data (taxable amounts for each month in 2021). In the experiments, we employed two different LSTM models: the traditional LSTM model and the improved LSTM model, to compare their performance.

The input dimension for the experiments was set to 1 to represent a single feature (taxable amount), and the hidden dimension (i.e., the number of hidden units) was set to 50. The loss function used was MSE, and the Adam algorithm was selected to optimize the model's weights and parameters. For the traditional LSTM model, the input consisted of data from corresponding months in 2019 to 2020 (e.g., taxable amounts for January 2019 and 2020 to predict data for January 2021). For the improved LSTM model, the input dimension, hidden dimension, loss function, and optimization algorithm remained consistent, but we introduced an additional set of interpolated data into the input, namely, the average of data from corresponding months in 2019 to 2020, resulting in 12 new data points. Consequently, we increased the number of input data points from 2 to 3, with the aim of smoothing the original data and reducing spikes and fluctuations to improve prediction accuracy.

4.2 Experimental Results

This study conducted experiments using real tax data to compare the performance of the traditional LSTM model with an enhanced LSTM model in tax data prediction. By analyzing the predictive outcomes of both models, several key findings were obtained. Firstly, we employed the traditional LSTM model to predict the tax data for corresponding months from 2019 to 2020, followed by a comparison with the actual taxable amounts for each month in 2021. We also performed the same prediction task using the improved LSTM model. The following are the main findings from the experimental results:

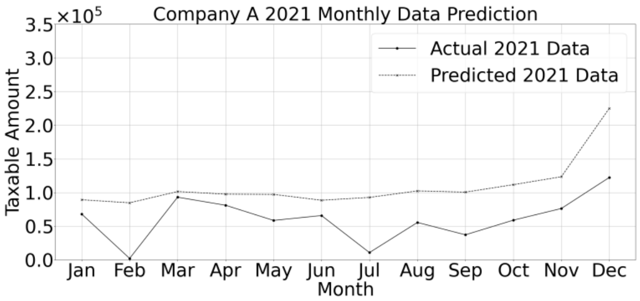


Figure. 3Comparison between LSTM Model Predictions and Actual Data for CompanyA

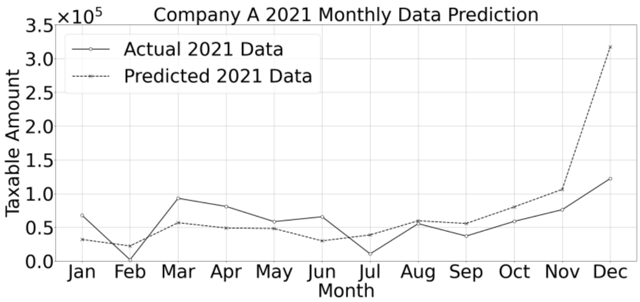


Figure. 4Comparison between Enhanced LSTM Model Predictions and Actual Data for Company A

For Company A, in Figure 3, we observe a comparison between the predictions of the traditional LSTM model and the actual data. It is evident that the general trend is similar, but there is a significant disparity in numerical values. In Figure 4, the enhanced LSTM model's predictions are compared to the actual data, showing an improvement in addressing the issue of substantial numerical disparities.

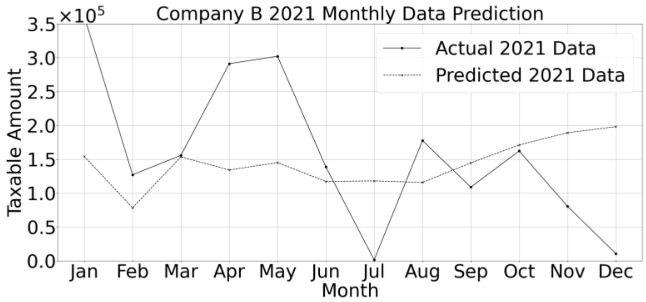


Figure. 5Comparison between LSTM Model Predictions and Actual Data for Company B

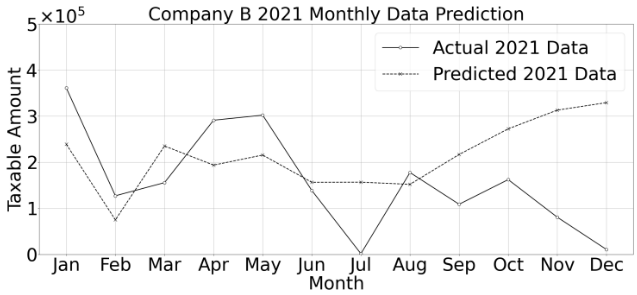


Figure. 6Comparison between LSTM Model Predictions and Actual Data for Company B

The above figures represent the forecast results based on tax data for Company B. In Figure 5, we can observe the comparison between the predictions of the traditional LSTM model and the actual data, especially noting a significant numerical disparity in the later months. Figure 6 shows the comparison between the predictions of the enhanced LSTM model and the actual data, demonstrating a relative improvement in addressing the issue of substantial numerical disparities between predicted and actual tax data.

4.3 Experimental Evaluation

In this section, we will present the performance evaluation results of our proposed enhanced LSTM model for tax revenue prediction. To assess the effectiveness of the model, we employed two commonly used performance evaluation metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). These metrics provide valuable insights into the accuracy and precision of the model's predictions.

MAE is a metric that measures the average magnitude of errors between model predictions and actual observations. It is calculated by averaging the absolute differences between the predicted values and the actual values for each data point. A lower MAE indicates that the model's predictions are closer to the actual observed values, suggesting that the model better captures underlying patterns in tax revenue data. The formula(1) for MAE is as follows:

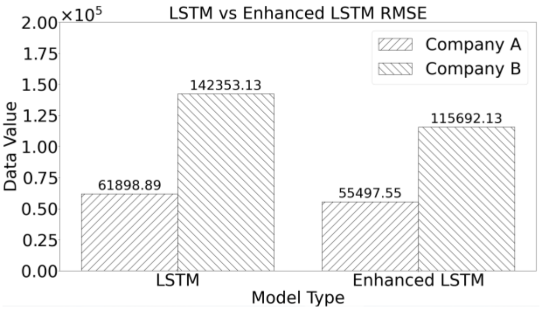
$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Y_{actual}\text{-}Y_{predicted}| \quad (1)$$

Where, n represents the total number of data points. Y_{actual} represents the actual tax revenue value. $Y_{predicted}$ represents the predicted tax revenue value.

RMSE is another widely used metric that takes into account the squared errors between predicted values and actual values, providing a measure of the magnitude of these errors. RMSE is particularly sensitive to large errors because it involves the square of the differences, penalizing significant deviations from actual values. The calculation formula(2) for RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |Y_{actual} - Y_{predicted}|} \quad (2)$$

Where, n represents the total number of data points. Y_{actual} represents the actual tax revenue value. $Y_{predicted}$ represents the predicted tax revenue value. Using MAE and RMSE as performance evaluation metrics allows for a reasonably accurate quantification of performance variations.



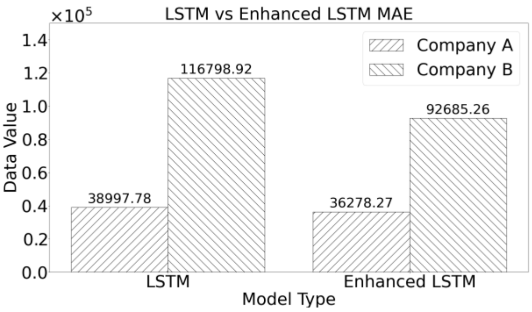


Figure. 7MAE and RMSE Metrics for Predictive Results

From Figure 7, we can observe the numerical comparison of performance evaluation metrics for the two companies under the two LSTM models. It is evident that under the improved LSTM model predictions, both MAE and RMSE show a decrease. In the case of Company B, the reduction in both metrics is more pronounced, and lower MAE and RMSE values indicate a better fit between predicted and actual data. Therefore, the enhanced LSTM model demonstrates superior predictive performance compared to the traditional LSTM model.

4.4 Advantages of the Enhanced LSTM Model

Based on the results and discussions from this experiment, we have identified several key advantages of the enhanced LSTM model compared to the traditional LSTM model when dealing with tax data. These improvements are primarily reflected in the following aspects:

Data Interpolation: The enhanced LSTM model incorporates data interpolation, which smoothes the raw data and reduces noise in the data, thereby enhancing the model's training and prediction capabilities. Data interpolation helps the improved model better capture the features of the data, especially when dealing with complex nonlinear relationships.

Better Fitting: By introducing data interpolation, the enhanced LSTM model can better fit the complex nonlinear relationships within tax data, leading to more accurate predictions of future tax trends. The improved model achieves a better fit to the data during the training process, reducing the impact of fluctuations and noise.

Performance Enhancement: Experimental results demonstrate a performance improvement of the enhanced LSTM model compared to the traditional LSTM model in tax data prediction. The average reductions in MAE and RMSE are 13.65% and 14.49%, respectively. This indicates that the improved model more accurately captures the characteristics and trends in tax data.

These advantages highlight the effectiveness of the enhanced LSTM model in enhancing predictive accuracy and handling complex tax data, making it a valuable tool for tax revenue forecasting.

5. CONCLUSION

In this study, we explored the application of traditional data analysis methods and machine learning in tax data analysis and prediction. We introduced traditional data analysis methods, including time series analysis, regression analysis, and hypothesis testing, along with their limitations when dealing with tax data. Subsequently, we delved into machine learning models, particularly the LSTM (Long Short-Term Memory) model, and its application in tax data prediction. We proposed an enhanced LSTM model that incorporates data interpolation to smooth raw data, reduce noise, and improve the model's training and prediction capabilities.

The experimental results demonstrate that the enhanced LSTM model outperforms the traditional LSTM model in tax data prediction tasks. Through data interpolation, the model better captures the nonlinear relationships within tax data, reduces prediction errors, and enhances predictive accuracy. The average reductions in Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are 13.65% and 14.49%, respectively, indicating the superiority of the improved model in tax revenue prediction.

6. ACKNOWLEDGEMENTS

We would like to express our heartfelt gratitude to all those who have contributed to this research. Firstly, we would like to thank our supervisor for his/her valuable guidance and insightful comments throughout the entire research process.

We also extend our sincere appreciation to all the participants who have generously devoted their time and effort to provide us with valuable data and insights. Finally, we would like to thank our families and friends for their unwavering support and encouragement. This research would not have been possible without their support. Also, we thank the lab funding from “Meteorological Information and Signal Processing Key Laboratory of Sichuan Higher Education Institutes of Chengdu University of Information Technology”; “Integrated Computing and Chip Security Sichuan Collaborative Innovation Center of Chengdu University of Information Technology”. The Open Project of National Intelligent Society Governance Testing Area (NO. ZNZL2023A04). the fund of the Scientific and Technological Activities for Overseas Students of Sichuan Province 2022(30) and funds from the Sichuan Provincial Department of Human Resources and Social Welfare” Researches on Key Issues of Edge Computing Server Deployment and Computing task Offloading”. Thank you for the funding, named "Research on the Construction and Management of Smart Civil Affairs Shared Data System" of Chengdu Civil Affairs Bureau. And Sichuan Science and Technology Program, Soft Science Project (No.2022JDR0076).

7. REFERENCES

1. Shu., Xiangbo, Zhang,.Liyan, Sun., Yunlian,Tang,,&Jinhui.(2021).Host-Parasite: Graph LSTM-in-LSTM for Group Activity Recognition.IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS,32(2),663-674.
2. Blanchet, Thomas, Flores,.Ignacio,Morgan., & Marc.(2022).The weight of the rich: improving surveys using tax data.JOURNAL OF ECONOMIC INEQUALITY,20(1),119-150.
3. Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48-62.
4. Vaupel, Y., Hamacher, N. C., Caspari, A., Mhamdi, A., Kevrekidis, I. G., &Mitsos, A. (2020). Accelerating nonlinear model predictive control through machine learning. *Journal of process control*, *92*, 261-270.
5. Szabó, B., &Babuška, I. (2021). Finite Element Analysis: Method, Verification and Validation.
6. Qi, J., Du, J., Siniscalchi, S. M., Ma, X., & Lee, C. H. (2020). On mean absolute error for deep neural network based vector-to-vector regression. *IEEE Signal Processing Letters*, *27*, 1485-1489.
7. Ćalasan, M., Aleem, S. H. A., &Zobaa, A. F. (2020). On the root mean square error (RMSE) calculation for parameter estimation of photovoltaic models: A novel exact analytical solution based on Lambert W function. *Energy conversion and management*, *210*, 112716.
8. Madeira, F., Pearce, M., Tivey, A. R., Basutkar, P., Lee, J., Edbali, O., ... & Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic acids research*, *50*(W1), W276-W279.
9. Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39-42.
10. Seiffert, D. J., McCarthy Veach, P., LeRoy, B., Guan, W., &Zierhut, H. (2019). Beyond medical actionability: Public perceptions of important actions in response to hypothetical genetic testing results. *Journal of Genetic Counseling*, *28*(2), 355-366.
11. Charbuty, B., &Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20-28.
12. Probst, P., Wright, M. N., &Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301.
13. Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189-215.
14. Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*, *31*(7), 1235-1270.
15. Yang, Y., Li, L., Lin, R. B., Ye, Y., Yao, Z., Yang, L., ... & Chen, B. (2021). Ethylene/ethane separation in a stable hydrogen-bonded organic framework through a gating mechanism. *Nature Chemistry*, *13*(10), 933-939.
16. Wang, B., Zhang, N., Lu, W., & Wang, J. (2019). Deep-learning-based seismic data interpolation: A preliminary result. *Geophysics*, *84*(1), V11-V20.