Predicting Socioeconomic Inequality Using Machine Learning: A Study on Multi-Source Big Data Fusion Model

Guoce Shui *

Guangzhou Foreign Language School, Guangzhou, 511455, China

shui45693@gmail.com

* Corresponding author

Abstract

Socioeconomic inequality hinders sustainable development, and its prediction is of great significance for formulating governance policies. Current studies have limitations such as single-source data and models that struggle to capture temporal dependencies. This study focuses on county-level administrative units, integrates four types of data—non-target government affairs data, social media data, remote sensing data, and mobile payment data—to construct a dataset with 776 valid records, using the county-level Gini coefficient as the target variable. After data preprocessing and feature selection, an Attention-LSTM model is built for prediction. Experimental results show that the Attention-LSTM model performs well, with a Root Mean Square Error (RMSE) of 0.021, a Mean Absolute Percentage Error (MAPE) of only 7.7131%, and a coefficient of determination (R^2) of 0.8497 on the test set, which is significantly better than traditional machine learning models such as XGBoost, LSTM, and Random Forest (RF). The multi-source data fusion framework and Attention-LSTM model constructed in this study provide a technical means for monitoring socioeconomic inequality at the county level, helping policymakers identify key intervention areas and formulate differentiated governance policies.

Keywords

Socioeconomic Inequality, Multi-Source Data, Attention-LSTM, Gini Coefficient.

1. Introduction

Inequality, defined as the uneven distribution of income, consumption, wealth, or opportunities among different groups within a society, has long been regarded as an obstacle to sustainable development [1]. The World Bank uses the Gini coefficient as a standard to measure the degree of inequality and classifies countries with a Gini coefficient exceeding 0.4 as "highly unequal countries". According to its report released in 2024, 52 countries worldwide have been included in the category of highly unequal countries. A high level of inequality can hinder poverty reduction efforts, inhibit economic growth, limit individuals' access to economic and educational opportunities, and weaken the overall social cohesion within a country. Conversely, effectively alleviating inequality can inject important impetus into high-quality economic development and the improvement of human capital levelst [2].

Research on the prediction of socioeconomic inequality has achieved certain results both domestically and internationally, but there are also many shortcomings. At the data level, most current studies mainly rely on traditional single statistical data sources, such as census data and economic statistics, while the utilization of new types of data, such as social media data and mobile device location data, is relatively limitedt [3, 4]. These new types of data can reflect socioeconomic phenomena from more dimensions, and the lack of integration makes the portrayal of socioeconomic inequality less comprehensive and in-depth. At the model level,

traditional machine learning algorithms are commonly used in current research on socioeconomic inequality prediction. Although these algorithms can handle some nonlinear relationships, they struggle to effectively capture long-term dependencies in socioeconomic data with temporal characteristics, thereby affecting the accuracy and forward-looking nature of predictionst [5, 6]. While deep learning models have shown strong capabilities in some fields, their application in socioeconomic inequality prediction research is mostly limited to single-data scenariost [7]. Existing studies have not yet formed a framework that combines multi-source data with deep learning, failing to fully utilize the complementarity between multi-source data and the ability of deep learning to fit complex relationships.

Based on this, this study takes county-level administrative units as the research object, systematically integrates four types of multi-source feature variables—non-target government affairs data, social media data, remote sensing data, and mobile payment data—to construct a multi-dimensional feature dataset. Then, using the internationally accepted Gini coefficient as the target variable, an Attention-LSTM model is created to conduct research on socioeconomic inequality prediction. This study aims to provide technical support for the prediction of socioeconomic inequality and at the same time offer references for formulating more targeted inequality governance policies.

2. Methodology

2.1. Dataset

The socioeconomic inequality dataset used in this study takes county-level regions as the research carrier, containing a total of 776 valid observation records and covering 14 core variables. According to the variable functions and research objectives, the dataset can be divided into two categories: "input feature variables" and "target variable". Among them, the input feature variables are further subdivided into non-target government affairs variables, social media variables, remote sensing variables, and mobile payment variables, as shown in Table 1.

Table 1: Variables in the Dataset

Variable Type	English Full Name	English Abbreviation
Target Variable	County-Level Gini Coefficient	CLGC
Government Variable	Urban Unemployment Rate	UUR
	Proportion of Education Expenditure in Fiscal Expenditure	PEEFE
	Number of Beds in Medical and Health Institutions	NBMIHI
	Coverage Rate of Urban Employee Endowment Insurance	CRUEEI
Social Media Variable	Frequency of Employment Anxiety Keywords	FEAK
	Frequency of Consumption Complaint Keywords	FCCK
	Frequency of Housing Pressure Keywords	FHPK
	Frequency of Educational Resource Anxiety Keywords	FERAK
Remote Sensing Variable	Distribution Entropy of Public Service Facilities	DEPSF
	Average Nighttime Light Intensity	ANLI
	Proportion of Built-Up Area	PBUA
	Green Space Coverage Rate	GSCR
Mobile Payment Variable	Proportion of Payment Frequency of High-Consumption Groups	PFHCG
	Proportion of Payment for Livelihood Consumption	PPLC
	Proportion of Cross-Regional Payment Amount	PCRPA

(1) Target Variable: County-Level Gini Coefficient (CLGC): An indicator to measure the fairness of income distribution among residents at the county level (with a value range of 0-1). A higher value indicates a larger income gap.

(2) Non-Target Government Affairs Variables: Including UUR, PEEFE, NBMIHI, and CRUEEI. UUR reflects the unemployment situation of the urban labor force, calculated as the percentage of the number of urban unemployed people in the total labor force; a higher value indicates a more prominent unemployment problem. PEEFE reflects the government's investment in education, i.e., the proportion of educational expenditure in total fiscal expenditure; a higher proportion indicates greater emphasis on education. NBMIHI measures the regional medical hardware supply capacity, referring to the total number of beds in all local medical and health institutions; a larger number means more sufficient medical resources. CRUEEI reflects the popularity of urban employees' participation in endowment insurance; a higher coverage rate indicates a more complete social security system.

(3) Social Media Variables: Including FEAK, FCCK, FHPK, and FERAK. FEAK counts the number of occurrences of keywords related to employment anxiety on social media; a higher frequency indicates a stronger sense of employment anxiety among the public. FCCK records the number of occurrences of consumption-related complaint keywords on social media; a higher frequency indicates obvious dissatisfaction with consumption among the public. FHPK counts the number of occurrences of keywords related to housing pressure on social media; a higher frequency indicates a strong perception of housing pressure among the public. FERAK records the number of occurrences of keywords related to anxiety about educational resources on social media; a higher frequency reflects a strong sense of anxiety among the public regarding the distribution of educational resources.

(4) Remote Sensing Variables: Including DEPSF, ANLI, PBUA, and GSCR. DEPSF measures the balance of the distribution of regional public service facilities (such as schools and hospitals); the closer the entropy value is to the theoretical optimal value, the more balanced the distribution. ANLI reflects the level of regional economic activity and population agglomeration; a higher average value usually represents a higher level of regional development. PBUA refers to the proportion of the built-up area in the total regional area; a higher proportion indicates a higher level of urbanization in the region. GSCR measures the proportion of green space area in the total area; a higher coverage rate represents better ecological environment quality.

(5) Mobile Payment Variables: Including PFHCG, PPLC, and PCRPA. PFHCG calculates the proportion of the mobile payment frequency of high-consumption groups in the total payment frequency; a higher proportion indicates higher activity of high-consumption groups. PPLC refers to the proportion of the mobile payment amount for livelihood consumption (such as food and medical care) in the total payment amount; a higher proportion indicates strong demand for livelihood consumption. PCRPA counts the proportion of cross-regional mobile payment amount in the total payment amount; a higher proportion indicates more frequent economic exchanges between regions.

2.2. Data Preprocessing

2.2.1. Missing Value Handling

In the process of data preprocessing, handling missing values is an important link to ensure subsequent analysis. Due to differences in the generation mechanisms, distribution characteristics, and missing situations of different types of data, differentiated imputation strategies are adopted, as follows:

For non-target government affairs variables (UUR, PEEFE, NBMIHI, CRUEEI): This type of data is derived from administrative statistics, with high standardization and a missing rate of less than 5%. The "intra-provincial mean imputation method" is adopted. Taking the province where the target county is located as the scope, the mean value of the corresponding variable

among all valid observation counties in the province is calculated and directly used as the result of missing value imputation.

For mobile payment variables (PFHCG, PPLC, PCRPA): Missing values are related to commercial activity. The "intra-prefecture-level city mean imputation method" is adopted. Taking the prefecture-level city where the target county is located as the scope, the mean value of the corresponding variable among valid counties within the prefecture-level city is calculated and used as the result of missing value imputation.

2.2.2. Outlier Handling

In this study, the " 3σ criterion" is used to identify and handle outliers for all input feature variables. First, the threshold for outlier determination is determined through statistical calculation, and then the identified outliers are replaced. A variable value of a sample is determined as an outlier if it satisfies:

$$|x_i - \mu| > 3\sigma$$

In the formula, x_i is the variable value of the i-th sample, μ is the global mean of the variable, and σ is the global standard deviation of the variable. Finally, the outliers are processed by replacing them with the global mean of the variable.

2.2.3. Standardization

To eliminate the impact of variable dimension differences on model training, Z-score standardization is used to process all input feature variables. The formula is as follows:

$$x' = \frac{x - \mu}{\sigma}$$

Among them, x is the original value of the variable, μ is the global mean of the variable, and σ is the global standard deviation of the variable. After standardization, all input feature variables have a mean of 0 and a standard deviation of 1.

2.3. Feature Selection

To identify the linear dependencies between original variables and investigate the interference of multicollinearity on subsequent modeling, this study uses the Pearson correlation coefficient to construct a fullvariable correlation matrix and quantitatively analyze the strength of the linear correlation between each input feature. The formula for the Pearson correlation coefficient is as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - x)(y_i - y)}{\sqrt{\sum_{i=1}^{n} (x_i - x)^2}\sqrt{\sum_{i=1}^{n} (y_i - y)^2}}$$

Among them, r_{xy} is the correlation coefficient between feature x and feature y, with a value range of $[-1, 1]$; x_i and y_i are the i -th sample values of the two features respectively; x and y are the sample means of the two features respectively; n is the total number of samples. A heatmap of the correlation coefficient matrix between features is drawn, and the results are shown in Figure 1.

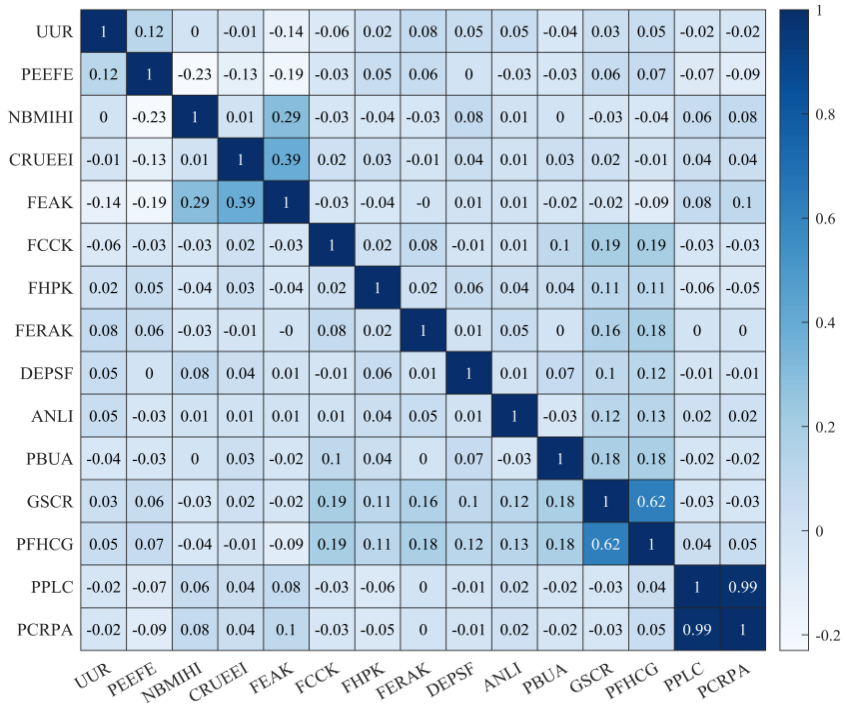

Figure 1: Feature Correlation Matrix

It can be seen from the figure that the correlation coefficient between feature GSCR and PFHCG reaches 0.62, showing a strong positive correlation; the correlation coefficient between feature PPLC and PCRPA is as high as 0.99, which is almost a perfect positive correlation. During feature selection, for feature combinations with high linear correlation, one of the more representative or more interpretable features is usually retained. Based on this, it is considered to eliminate PFHCG and PCRPA, and a total of 13 features are retained to reduce redundant information between features and improve the efficiency and accuracy of subsequent model training and prediction.

2.4. Attention-LSTM

This study takes county-level multi-source data as input and the county-level Gini coefficient as the prediction target. Considering that multi-source features have different contribution degrees to inequality prediction and it is necessary to capture potential nonlinear correlations between features, the traditional LSTM model is difficult to focus on key features. Therefore, the attention mechanism is introduced and combined with LSTM to construct the Attention-LSTM model. The model mainly includes an input layer, an attention layer, an LSTM layer, and a fully connected output layer. It not only retains the ability of LSTM to fit complex relationships but also strengthens the weight of core features through the attention layer, thereby improving prediction accuracy.

2.4.1. Attention Mechanism

A single-head attention mechanism is introduced to calculate the importance weight of each embedded feature to the target variable (CLGC), with the following steps:

Step 1: Generate query (Q), key (K), and value (V) matrices. Based on the embedded feature E, Q, K, and V are generated through three independent fully connected layers respectively, where:

$$Q = E \cdot W_q + b_q$$
$$K = E \cdot W_k + b_k$$
$$V = E \cdot W_v + b_v$$

W_q, W_k, W_v are weight matrices, $b_q, b_k, b_v \in \mathbb{R}^d$ are bias vectors, and the outputs $Q, K, V \in \mathbb{R}^{n \times d}$.
Step 2: Calculate attention scores and weights. Scaled Dot-Product Attention is used to calculate the scores to avoid excessively large inner product values caused by high dimension d. The formula is:

$$\text{Score}(Q, K) = \frac{Q \cdot K^T}{\sqrt{d}}$$

The score matrix is normalized by the softmax function to obtain the attention weight α of each feature. The formula is:

$$\alpha = \text{softmax}(\text{Score}(Q, K))$$

Among them, $\alpha \in \mathbb{R}^{n \times d}$, and $\sum_{j-1}^{d} \alpha_{ij} = 1$, where α_{ij} is the weight of the j -th feature of the i -th sample.
Step 3: Generate weighted features. Multiply the weight matrix by the value matrix V to obtain the attentionweighted feature matrix A. The formula is:

$$A = \alpha \cdot V$$

2.4.2. LSTM Layer

The LSTM layer is used to capture nonlinear correlations and potential dependencies between weighted features. The LSTM unit controls the transmission and update of information through the forget gate, input gate, and output gate. The core formulas are as follows:
Forget gate: Determines the proportion of historical hidden states to be retained. The formula is:

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, A_t] + b_f\big)$$

Input gate: The proportion of new information used to update the cell state. The formulas are:

$$i_t = \sigma(W_i \cdot [h_{t-1}, A_t] + b_i)$$
$$\tilde{C}_t = \tanh\,(W_C \cdot [h_{t-1}, A_t] + b_C)$$

Cell state: Updates long-term memory information. The formula is:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

Output gate: Generates the current hidden state. The formulas are:

$$o_t = \sigma(W_o \cdot [h_{t-1}, A_t] + b_o)$$
$$h_t = o_t \odot \tanh\,(C_t)$$

Among them, $W_f, W_i, W_C, W_o \in \mathbb{R}^{(d+h) \times h}$ are the weight matrices of the LSTM layer, $b_f, b_i, b_C, b_o \in \mathbb{R}^h$ are bias vectors, σ is the Sigmoid activation function, tanh is the hyperbolic tangent activation function, $\odot$ is element-wise multiplication, and $h_t \in \mathbb{R}^{n \times h}$ is the hidden state matrix output by the LSTM layer.
The 776 valid samples are randomly divided into a training set, a validation set, and a test set in a ratio of 7:1.5:1.5. The training set is used for model parameter learning, the validation set

is used to adjust hyperparameters during iteration to avoid overfitting, and the test set is used to independently evaluate the final generalization performance of the model. In this study, the final set Learning rate is 0.005, Batch-size is 32 , and Max epochs is 100. In addition, a Dropout layer is added between the LSTM layer and the fully connected output layer to reduce the risk of model overfitting, and the Dropout rate is set to 0.2.

2.5. Model Evaluation Indicators

To comprehensively and accurately evaluate the performance of the Attention-LSTM model in the task of socioeconomic inequality prediction, this study selects three indicators: Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and coefficient of determination (R^2) to analyze the model.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\%$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Among them, n represents the number of samples in the test set, $\hat{y}_i$ is the predicted value of the Gini coefficient for the i -th sample in the test set, and y_i is the true value of the Gini coefficient for the i -th sample.

3. Results

It can be seen from the loss function change curve shown in Figure 2 that the Attention-LSTM model exhibits good convergence characteristics during the training process. In the early stage of training (1-200 Epochs), the model's training loss value decreases rapidly, indicating that the model can quickly learn the basic correlation between multi-source input features and the county-level Gini coefficient. As the number of training epochs increases (200-800 Epochs), the rate of decrease in the loss value gradually slows down, the curve tends to be flat, and the model parameters gradually approach the optimal solution. When the number of training epochs exceeds 800 Epochs, the loss value is basically stable. Combined with the change in the validation set loss, it can be judged that the model does not produce overfitting.

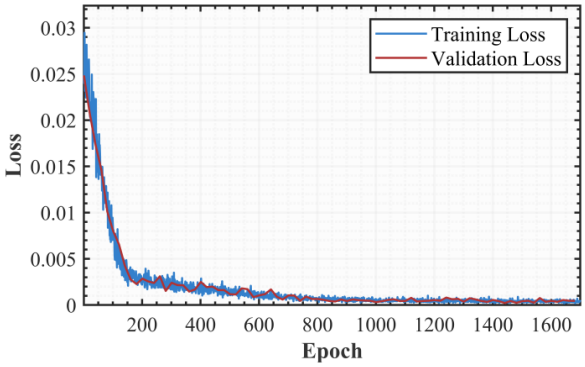

Figure 2: Loss Curve of Attention-LSTM model

From the prediction results of the Attention-LSTM model on the test set shown in Figure 3, the change trend of the predicted values is highly consistent with that of the true values. The deviation between the predicted values and the true values of most samples is small, and only slight deviations occur in a few samples. This intuitively indicates that the model can accurately capture the actual distribution characteristics of the county-level Gini coefficient and has good prediction ability for different levels of socioeconomic inequality.

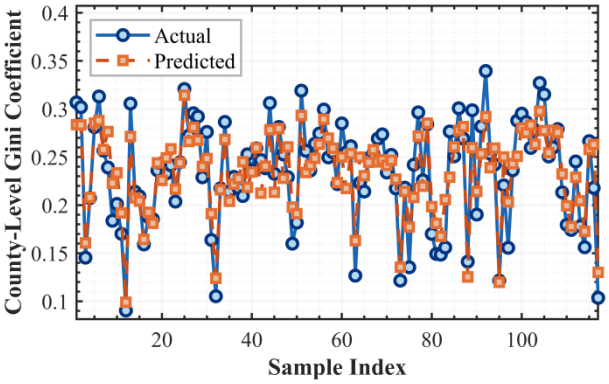

Figure 3: Prediction Results of the Attention-LSTM Model (Test Set)

Further, combined with the quantitative index comparison of the prediction results of various models in Table 2, it can be seen that the Attention-LSTM model is significantly better than other traditional machine learning models and the single LSTM model in all evaluation indicators. The RMSE of the Attention-LSTM model is 0.021, which is 26.57%, 46.58%, 51.16%, and 63.73% lower than that of XGBoost (0.0286), LSTM (0.0395), SVR (0.0432), and MLR (0.0579) respectively. This indicates that the model has lower prediction deviation and lower sensitivity to extreme values, and better output stability. The MAPE is only 7.7131%, which is lower than that of other comparison models, and the average percentage deviation between the predicted values and the true values is less than 8%. The R^2 reaches 0.8497, which is higher than that of XGBoost (0.8106), LSTM (0.8025), etc. The explanation rate for the variation of the county-level Gini coefficient is about 85%, indicating that the model has relatively reliable explanatory power for the driving mechanism of socioeconomic inequality and reliable prediction results.

Table 2: Comparison of Prediction Results of Various Models

Model	RMSE	MAPE (%)	R2
Attention - LSTM	0.021	7.7131	0.8497
XGBoost	0.0286	8.5927	0.8106
LSTM	0.0395	10.8452	0.8025
RF	0.0358	9.4168	0.7843
SVR	0.0432	11.6739	0.7579
MLR	0.0579	14.9826	0.7318

4. Conclusions

This study takes county-level administrative units as the research object, constructs a multi-source feature dataset by integrating non-target government affairs data, social media data, remote sensing data, and mobile payment data, and conducts research on socioeconomic inequality prediction based on the Attention-LSTM model. The core results are as follows: At the data level, the integration of multi-source data breaks through the limitations of traditional single statistical data. After feature selection to eliminate highly correlated variables, 13 core features are retained. At the model level, the Attention-LSTM model strengthens the weight of core features through the attention mechanism and captures the temporal dependencies of

data through the LSTM layer. It achieves prediction performance with RMSE=0.021, MAPE=7.7131%, and R^2=0.8497 on the test set, which is significantly better than traditional models such as XGBoost, LSTM, and RF, and has high accuracy and strong generalization ability. The research results provide technical support and practical references for the governance of socioeconomic inequality. The constructed multi-source data fusion framework and Attention-LSTM model can provide a reliable tool for the dynamic monitoring and early warning of socioeconomic inequality at the county level. In addition, the model can help policymakers identify key intervention areas, and its identification of core features can guide the formulation of differentiated policies, providing data-driven decision-making basis for alleviating socioeconomic inequality and promoting coordinated regional development.

References

[1] Arya P K, Sur K, Dhote S, et al. Integrating Multi-Source Satellite Imagery and Socio-Economic Household Data for Wealth-Based Poverty Assessment of India: A GIS and Machine Learning Based Approach: Arya et al [J]. Social Indicators Research, 2025: 1-24.

[2] Niu T, Chen Y, Yuan Y. Measuring urban poverty using multi-source data and a random forest algorithm: A case study in Guangzhou [J]. Sustainable Cities and Society, 2020, 54: 102014.

[3] Zhao X, Yu B, Liu Y, et al. Estimation of poverty using random forest regression with multi-source data: A case study in Bangladesh [J]. Remote Sensing, 2019, 11(4): 375.

[4] Ji J. Machine Learning-Based Income Inequality Prediction: A Case Study [C]//Proceedings of the 2024 2nd International Conference on Artificial Intelligence, Systems and Network Security. 2024: 34-39.

[5] Fan C, Xu J, Natarajan B Y, et al. Interpretable machine learning learns complex interactions of urban features to understand socio-economic inequality [J]. Computer-Aided Civil and Infrastructure Engineering, 2023, 38(14): 2013-2029.

[6] Reza S A, Rahman M K, Hossain M S, et al. AI-Driven Socioeconomic Modeling: Income Prediction and Disparity Detection Among US Citizens Using Machine Learning [J]. Advances in Consumer Research, 2025, 2(4).

[7] Pradhan N, Agrawal A. Mapping fine-scale socioeconomic inequality using machine learning and remotely sensed data [J]. PNAS nexus, 2025, 4(2): pgaf040.