Data-Driven Approaches for Accurate Temperature Forecasting: A Machine Learning Perspective

Juntao Chen.

School of Mathematics and Big Data, Chongqing University of Arts and Sciences

Juntaochen1@126.com

Abstract

This study delves into data-driven methods for predicting temperatures, viewed through the lens of machine learning. Three regression models - Random Forest, Decision Tree Regression, and Multiple Linear Regression were rigorously compared using real-world datasets. The results underscore the superiority of Random Forest, boasting an impressive R^2 score of 0.9831, outperforming Decision Tree Regression (0.9732) and Multiple Linear Regression (0.9297) in terms of predictive accuracy. Notably,Random Forest also demonstrated the lowest MAE and MSE values, underscoring its precision in temperature prediction. Overall, this study highlights Random Forest as a highly promising model, advancing machine learning techniques for temperature forecasting and offering valuable insights for practical applications.

Keywords: Temperature Prediction; Variable Selection; Rando m Forest； Tree Regression； Multiple Linear Regression

Methodology

Firstly, the data preprocessing and Featureselection temperature data series based on PCC are studied. On this basis, three machine learning prediction models were constructed to achieve accurate temperature prediction. Based on the estimation of point prediction error, compare and analyze the three models.

1. Pearson Correlation Coefficient

The linear correlation between two variables can be calculated and measured using the traditional PCC approach. PCC has a value between -1 and 1; the more significant the absolute value, the stronger the correlation. The PCC between X and Y is shown in Equation (1).

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{D_X}\sqrt{D_Y}} \quad (1)$$

2. Multiple Linear Regression

Multiple linear regression analysis is used to analyze the linear relationship between a single dependent variable and multiple independent variables. It assesses the presence of multicollinearity among the dependent variable and independent variables based on tolerance and variance inflation factor. The general form of multiple linear regression can be expressed as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon. \quad (2)$$

3. Decision Tree Regression

Decision tree learning divides data using attribute values,e.g., ID3, C4.5, CART. C4.5 excels in efficiency and handling missing data, using info gain ratio for attribute selection. In the formula (3), signifies info gain ratio with attribute A as the split node.

$$R_g(D,A) = \frac{H(D) - H(D|A)}{H_A(D)}. \quad (3)$$

4. Random Forest

Random Forest (RF) is a highly effective method for predicting temperature. By leveraging its ensemble of decision trees, RF captures the complex relationships and patterns in temperature data. It overcomes the limitations of traditional statistical and physical models by directly learning from the data.

Results and discussion

★ Data description and preprocessing

By undergoing data processing, the dataset emerges asrefined, cohesive, and primed for analysis (Figure 1).This serves as the bedrock for subsequent modeling and prediction phases, empowering diverse algorithms to craft precise and dependable temperature prediction models.

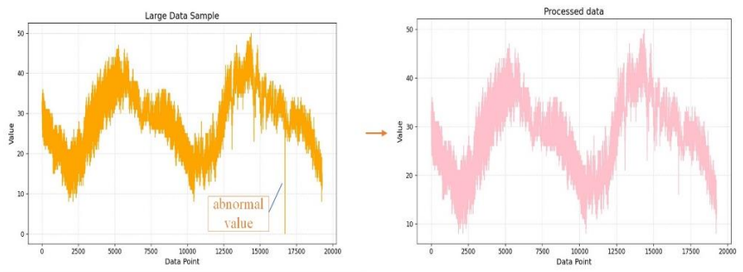

Figure 1. Comparison chart before and after data processing

★ Feature selection results

In the context of temperature prediction, Feature selection plays a crucial role in identifying the most relevant variables that affect temperature fluctuations. In this study, we used Pearson correlation coefficient (PCC) as a measure to evaluate the linear relationship between each feature and the target variable. The selected results are shown in Figure 2, and the visualization of the processed data is shown in Fig. 3.

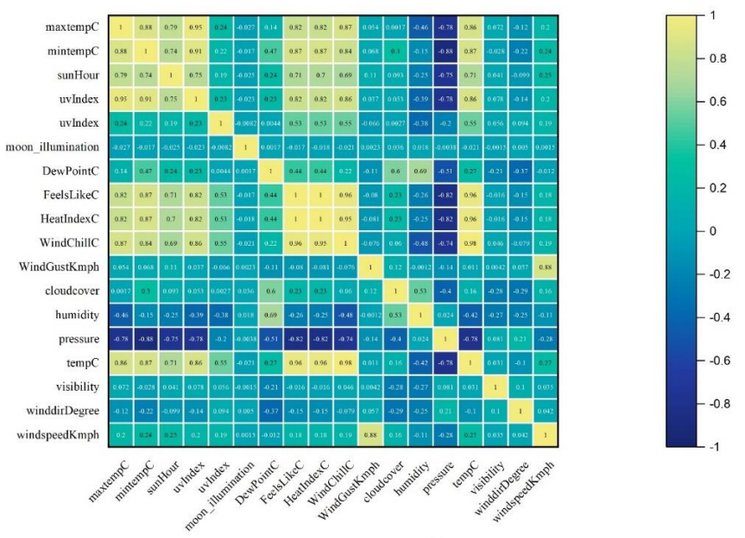

Figure 2. Correlation coefficient diagram

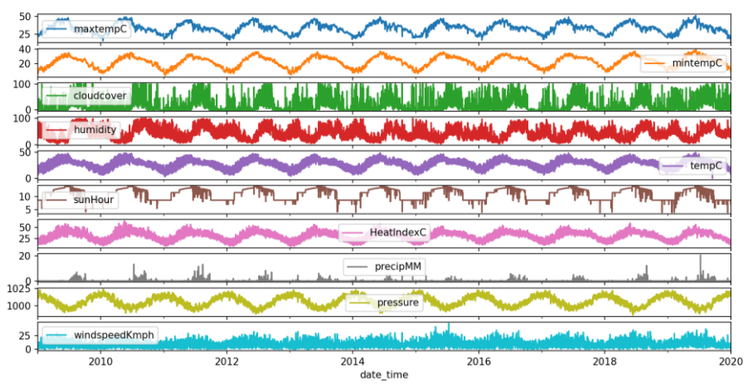

Figure 3. Graph of each factor from 2009 to 2020

★ Multiple linear regression prediction

The multiple linear regression model used for temperature prediction yielded promising results, as indicated by the evaluation metrics. The results are shown in Figure 4.

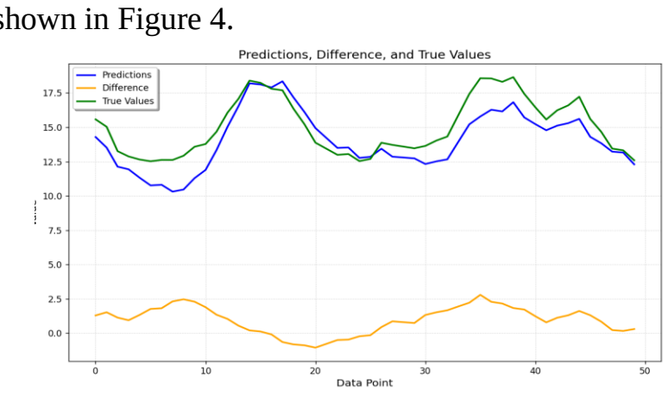

Figure 4. Local comparison between predicted and true values .

★ Decision Tree Regression Prediction

The decision tree regression model employed for temperature prediction exhibited excellent performance, as indicated by the evaluation metrics. we took a comparison graph of the last 50 steps, as shown in Figure 5.

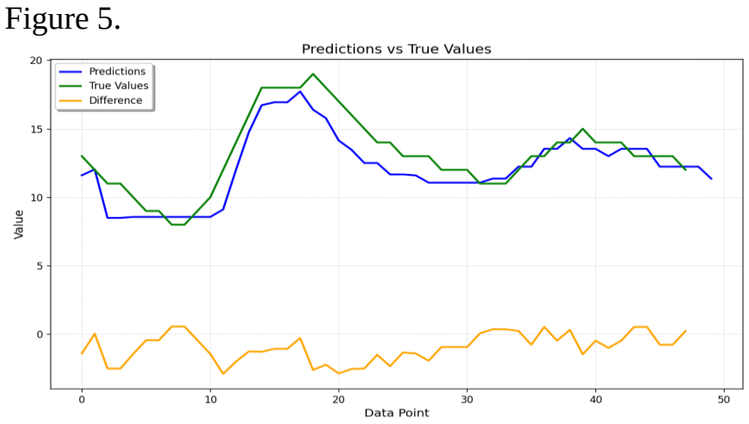

Figure 5 Local comparison between predicted and true values

★ Random forest prediction

The random forest model utilized for temperature prediction demonstrated exceptional performance, as indicated by the evaluation metrics. The results are shown in Figure 6.

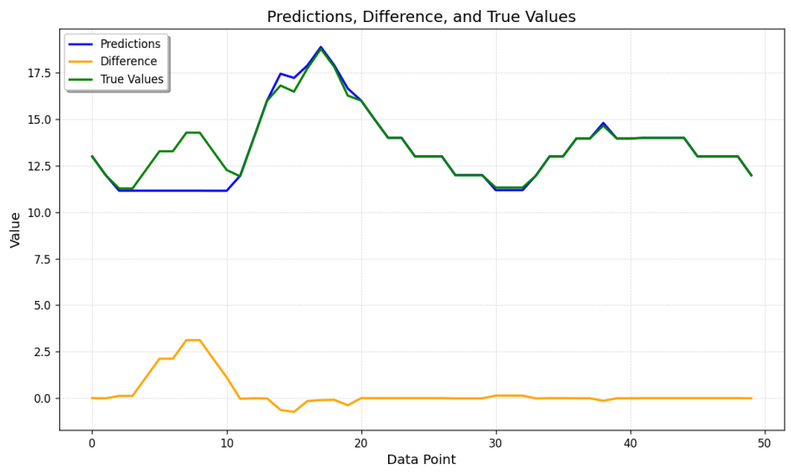

Figure 6. Local comparison between predicted and true values

Table 1. Comparison of Results of Three Models

Model	MAE	MSE	R^2
Multiple linear regression	1.11323725	1.87309111	0.9297275
Decision Tree Regression	0.53767023	0.714699443	0.9731867
Random forest	0.46911480	0.446849133	0.9832356

Conclusion

In this study, we are committed to solving the problem of temperature prediction and have drawn some important conclusions by collecting and analyzing relevant data. First, we establish a temperature prediction model based on random forest regression, and use historical meteorological data as input characteristics. Through training and testing of the model, we found that it can accurately predict the trend and fluctuation of temperature changes.

Reference

[1] Xu Kangkang, et al. "Cement rotary kiln temperature prediction based on time-delay calculation and residual network and bidirectional novel gated recurrent unit multi-model fusion." Measurement 218.

[2] Jin, C. (2021). Comparative study of the prediction of high-temperature flow stress of AZ80 magnesium alloy using physical constitutive models and BP artificial neural network models. Rare Metal Materials and Engineering, 50(11), 3924-3933..

[3] Wang, Y., et al. (2021). A statistical downscaling method for regional seasonal climate prediction based on global dynamical models and the SMART principle. Meteorological Science, 41(05), 569-583.

Due to space limitations, only three references are included.