Research Analysis of Optimization of Temperature Prediction Model Based on Random Forest

Tianhua Jiang

Zhejiang University of Finance and Economics Hangzhou, Zhejiang Province, China
1726499161@zufe.edu.cn

Abstract. Temperature prediction plays a crucial role in meteorological analysis, scientific planning of agricultural production, precise allocation of energy management and so on. But traditional temperature prediction methods often fall short in terms of prediction accuracy and fail to meet expectations. To address this challenging issue, this study innovatively proposes a new temperature prediction approach based on in-depth improvement of the random forest model. High-quality feature sets rich in information are constructed through data preprocessing procedures and feature engineering. The optimal parameter configuration is explored using a hyperparameter tuning strategy, and the improved random forest regression model is employed to achieve accurate temperature prediction. Outlier removal, in-depth analysis of feature importance, and rigorous considerations of cross-validation are integrated to enhance the robustness of the model. On the independent test set, this method demonstrates remarkable experimental results: the prediction accuracy reaches a high level of 94.22%, and the Root Mean Squared Error (RMSE) drops sharply to 4.85, showing a significant performance improvement compared with the baseline model. Undoubtedly, this study confirms that the optimized random forest model is highly effective and robust in the field of temperature prediction, thereby providing a reliable machine learning solution for in-depth analysis of meteorological data.

Keywords: Temperature Prediction; Random Forest; Hyperparameter Tuning; Machine Learning

1 Introduction

The accuracy of temperature prediction, which forms the cornerstone of the field of meteorological science, plays a vital role in the rational planning of agricultural production, efficient scheduling of energy systems, and effective early warning of natural disasters [1-3]. In the past, temperature prediction mostly relied on time series analysis and numerical weather prediction models, which were inadequate in capturing nonlinear relationships and complex feature interactions. With the rapid advancement of machine learning technology, data-driven prediction methods that have emerged in

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_51

the meteorological community have opened up a new and promising technical exploration path for solving temperature prediction problems [4, 5].

Some papers have analyzed the spatiotemporal distribution characteristics of extreme temperature events in the Yellow River Basin and their response patterns to circulation, simulated extreme temperature indices using SMLR and RF models, and predicted future extreme temperature indices using circulation indices. These studies reveal the patterns and cause of extreme climate events in the Yellow River Basin, providing a scientific basis for extreme temperature forecasting [6]. There is also an experiment that designs a dynamic modeling machine learning approach based on three characteristics of seasonal-scale climate: model stability, feature factor complexity, and statistical relationship nonlinearity, and then tests its predictive ability for seasonal temperature and precipitation across the country [7]. There are also studies that utilize the Geographically Random Forest machine learning model to predict daily maximum and minimum temperatures across various regions [8]. Some articles have also proposed a new method based on CNN, which combines different data fusion and dimensionality reduction processes to predict temperature [9].

Meteorological data is full of noise and has complex correlations among various features, making it difficult to achieve the desired prediction results by directly applying standard machine learning models [10]. Facing this challenge, this paper proposes an enhanced random forest temperature prediction method, which can effectively build a high-precision prediction model based on small-scale meteorological data via reasonable data preprocessing, feature engineering, and thorough hyperparameter searching. This novel approach not only enhances the accuracy of the prediction substantially but also yields the most influential factors in temperature variations through feature importance analysis, paving a new direction in the exploration of meteorological data.

2 Methodology

2.1 Dataset Description

In this study, a public temperature dataset named temps.csv is used, which could be found in kaggle and contains rich and diverse time-series records of meteorological features. Within the meteorological context, this dataset is precisely the functional resource that not only covers multiple meteorological dimensions such as temperature and time but also specifically designates the actual temperature value as the target prediction variable. The selected data has undergone a rigorous quality control process to ensure that the samples used for model training are both representative and highly reliable.

2.2 Data Preprocessing

Data preprocessing is a crucial and well-planned stage for ensuring the integrity of the original dataset. First, regarding missing values: for numerical features, the median is wisely selected for imputation; for categorical features, the mode is skillfully used for supplementation, thereby ensuring the comprehensiveness and completeness of the

data. When dealing with the problem of outliers, the advanced detection method of Interquartile Range (IQR) is adopted; values outside the range [Q1 - 1.5IQR, Q3 + 1.5IQR] are skillfully constrained within the boundaries. This strategy is actually inspired by the essence of robust statistical methods, thus effectively reducing noise interference while subtly preserving the original distribution of the data, and further laying a more solid foundation for subsequent model training.

2.3 Feature Engineering

This study focuses on the goal of in-depth exploration of feature potential. Based on the consideration of achieving this goal, a comprehensive and detailed feature engineering process including multiple operation links (from basic feature extraction to complex feature construction) is carefully and systematically implemented, aiming to maximize the potential value of features in subsequent research and applications. This entire process is carried out on the basis of fully considering and balancing the overall research direction and expected results. Through the sophisticated technique known as one-hot encoding, categorical variables are transformed from their original state to a numerical form—this transformation is essentially an ingenious conversion from categorical to numerical form—with the purpose of enabling the model to process them without much difficulty. After the completion of this encoding operation, through a certain logic and processing procedure, the dimension of the original features is expanded in a very significant manner. These actions and the processed results further inject key elements of more sufficient input information (that meets specific conditions in a certain sense) into the subsequent operation process of the relevant model through specific logic and operations. Through the meticulous operation process of feature engineering, not only is the expressive ability of features significantly improved, but it also plays a supporting role in making it possible to capture the complex and changeable nonlinear relationships between features. All of this lays a solid foundation for the steady advancement of subsequent model training, although some key details still seem to be hidden in the fog and need further exploration.

2.4 Model Construction and Optimization

Random Forest Model. The random forest model, regarded as an integrated learning paradigm, is applied in related fields by carefully constructing a large number of decision trees and skillfully integrating their prediction insights, thereby effectively reducing the risk of overfitting and significantly enhancing the generalization ability of the model. In the exploration process of this study, after careful consideration, the random forest regressor is determined as the core architecture. The essence of this regressor lies in carefully forging multiple weak learners through the Bootstrap sampling strategy, and then ensuring the stability and reliability of its output by intelligently averaging the prediction results of these weak learners.

Hyperparameter Optimization Strategy. Given the problem that the random forest model is extremely sensitive to parameters, this study designs a manual hyperparameter tuning scheme in an innovative manner. During this process, eight different parameter

combinations are carefully tested, and at the same time, a series of parameters regarded as key are carefully adjusted and optimized.

n_estimators: The number of decision trees, which affects model complexity and stability.

max_depth: The maximum depth of the tree, which controls model complexity and the risk of overfitting.

min_samples_split: The minimum number of samples required to split an internal node.

min_samples_leaf: The minimum number of samples required for a leaf node.

To ensure the scientificity and reliability of parameter selection, a rigorous three-fold cross-validation method is adopted after careful consideration. This method is used to conduct a comprehensive evaluation of the performance of each parameter combination, in which the Mean Absolute Error (MAE) is used as the core evaluation metric.

Model Evaluation Metrics. To comprehensively evaluate model performance, this study adopts a multi-dimensional evaluation system:

Root Mean Squared Error (RMSE): Measures the degree of deviation between predicted values and actual values.

Mean Absolute Error (MAE): Evaluates the absolute magnitude of prediction errors.

Coefficient of Determination (R^2): Reflects the model's ability to explain the variation in the data.

Mean Absolute Percentage Error (MAPE): Provides a relative measure of errors.

3 Experimental Results

3.1 Model Performance Comparison

This study conducts an in-depth comparison of the prediction performance of three models: a simple baseline model (mean prediction), an unoptimized basic random forest model, and the carefully improved random forest model. As shown in Table 1, the optimized random forest model takes the lead in all evaluation metrics, demonstrating excellent performance.

Table 1. Comparison of Model Performance Results.

Model	RMSE	MAE	R^2	Accuracy (%)
Baseline Model	11.9	9.40	-0.01	84.66
Basic Random Forest	4.87	3.64	0.83	94.20
Improved Random Forest	4.85	3.63	0.84	94.22

The analysis results show that compared with the simple baseline model, the improved random forest achieves a significant reduction in the RMSE metric, and the accuracy increases by nearly 10%. Compared with the basic random forest, its RMSE

further decreases, which undoubtedly verifies the remarkable effectiveness of the optimization strategy.

3.2 Feature Importance Analysis

The feature importance of the optimized model is analyzed in depth, and key factors that have a significant impact on temperature prediction are accurately identified. Table 2 lists the top five important features and their contribution to the prediction results.

Table 2. Ranking of Feature Importance.

Ranking	Feature Name	Importance Score
1	temp_1 (The temperature a day ago)	0.683
2	average (Historical average temperature)	0.216
3	friend (The predicted value of a simple model)	0.027
4	day (Date (1-31st))	0.023
5	temp_2(The temperature two days ago)	0.020

In the feature importance analysis, not only the accumulated domain knowledge is verified for its rationality, but also solid data support is provided for the wisdom of feature selection, thereby helping to build a more streamlined and highly effective prediction model architecture.

3.3 Overfitting Analysis

With the purpose of thoroughly analyzing the generalization of the model, this paper although cautiously, compares the performance of the training set (participated in training) and the test set (applied to verify efficiency), among many processes, this is not easy as it seems. The tuned random forest classifier gave a superb training accuracy of 97.71% and it also achieved accuracy of 94.22% on the test data.The overfitting gap between the training set and the test set is as small as 3.50%. This situation seems to demonstrate the extremely excellent generalization ability of the model, and at the same time, there is no significant overfitting problem that could cause concern in its overall performance.

4 Conclusion

Focusing on the core issue of temperature prediction, this study innovatively proposes a prediction strategy based on a fully improved random forest. Through a series of operations including well-planned data preprocessing, feature engineering, and hyperparameter tuning, the paper has successfully constructed a high-precision temperature prediction model (by making full use of these methods). According to the experimental data, the prediction accuracy of this method on the test set is as high as 94.22%, which represents a significant improvement compared with traditional methods. The main contributions of this study are as follows:

In terms of data preprocessing, through a carefully designed and highly effective method, tricky problems in meteorological data such as outliers and missing values have been properly handled.

Guided by cross-validation, this study innovatively proposes a manual hyperparameter tuning strategy to accurately explore the parameter configuration suitable for the random forest model.

Through the analysis of feature importance (as you know), the key factors affecting temperature prediction have been identified as expected, thus clearly pointing out the direction for feature selection.

This prediction method is characterized by simplified parameters, which results in high training efficiency and stable prediction accuracy. It can be quickly deployed on edge devices with limited computing resources while maintaining stable prediction accuracy, making it particularly suitable for real-time control scenarios.

The random forest model designed in this article has achieved good results in temperature prediction, but there are still certain shortcomings and room for improvement, such as the lack of more meteorological data for training and the lack of certain predictions for occasional extreme weather

Therefore, meteorological data prediction will be promoted from the following two aspects in the future:

(1) Add other climate data to train and predict. For example, air pressure, particle size, humidity, precipitation, and geographic location are used to obtain more accurate predictions.

(2) Use and optimize merging other models to improve prediction accuracy.

References

1. Wu, Y., Cao, Z. L., Liu, C., et al.: Temperature field and curing degree prediction of large composite blades based on coupled finite element analysis and machine learning. Polymer Composites, (2025)
2. Wei, S., Wang, C., Zhang, F., et al.: Prediction of seasonal sea surface temperature based on temperature and salinity of subsurface ocean using machine learning. International Journal of Climatology, 44(5): 1326-1338. (2024)
3. Laukkarinen, A., Vinha, J.: Long-term prediction of hourly indoor air temperature using machine learning. Energy and Buildings, 325: 114972. (2024)
4. Wang, Y., Xu, Y., Song, X., et al.: Novel method for temperature prediction in rotary kiln process through machine learning and CFD. Powder Technology, 439: 119649. (2024)
5. Liu, C., Lu, Y., Feng, J., et al.: Prediction and customized design of Curie temperature of Fe-based amorphous alloys based on interpretable machine learning. Materials Today Communications, 38: 107667. (2024)
6. Chen, J. Q., Li, Y., Wang, B. et al.: Prediction of Extreme Temperature Events in the Yellow River Basin Based on Stepwise Multiple Linear Regression and Random Forest Model. Journal of Natural Disasters, 33(01):74-88. (2024) DOI: 10.13577/j.jnd.2024.0107.
7. Liu, J. N.: Research on Short term Climate Prediction in China Based on Machine Learning Methods. Nanjing University of Information Science and Technology. (2023) DOI: 10.27248/d.cnki.gnjqc.2023.000499.

8. Sailaja, B., Gayatri, S., Rathod, S., et al.: Spatial temperature prediction—a machine learning and GIS perspective. Theoretical and Applied Climatology, 155(11): 9619-9642. (2024)
9. Fister, D., Pérez-Aracil, J., Peláez-Rodríguez, C., et al.: Accurate long-term air temperature prediction with machine learning models and data reduction techniques. Applied Soft Computing, 136: 110118. (2023)
10. Rahman, S., Olausson, M., Vitucci, C., et al.: Machine learning-based ambient temperature prediction in radio access network environments. International Journal on Software Tools for Technology Transfer, 1-12. (2025)