EE&AE
2025 10th International Conference
on Energy Efficiency and Agricultural Engineering
5-7 November 2025, Starozagorski Bani, Bulgaria
IEEE

Regression Model Evaluation for Short-Term Temperature Prediction

Raluca-Alexandra Oană

Electronics, Telecommunication and Information Technology
National University of Science and Technology POLITEHNICA Bucharest,
raluca.oana@stud.etti.upb.ro

GOAL OF THE STUDY

The goal of this study is to compare the performance of Random Forest and Linear Regression models in predicting temperature values using a small dataset containing recorded temperatures over a three-year period.

METHODOLOGY OF THE INVESTIGATION

Before starting training a model it is necessary to consider other factors such as missing values, dependencies or data volume. The data used in this study consist of daily temperature records (minimum, maximum, and average) provided from 23 meteorological stations from Romania. Before training, a model is necessary to ensure that the input data is accurate and complete to obtain the best predictive behaviour. To achieve this stability the preprocessing part involved cleaning and splitting the data based on the geographical location and using feature engineering to extract separate features for the year, month and day to highlight the cyclical and seasonal pattern. All sets of data were preprocessed in the same way to maintain a fair comparison between Random Forest and Linear Regression. The principal characteristic of Supervised Learning algorithms is the manner the input data for the training is parted into training dataset and testing for the performance evaluation. This is why a model performance could be determined immediately after the training session by calculating the evaluation metrics Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and the Coefficient of Determination (R^2). For training the two models the data was split into training and testing data based on the model's complexity.

MAIN RESULTS FROM THE STUDY

An analysis of the error metrics during the training phase reveals that the Random Forest (RF) and Linear Regression (LR) models achieved very similar performance. For the mean temperature (TMED) data, the Random Forest model achieved an R^2 of 0.894, which is only fractionally higher than the Linear Regression's R^2 of 0.8879. This close performance is mirrored in the error metrics: the MSE for RF was 8.628 compared to LR's 9.1883, and the MAE was 2.232 versus LR's 2.3450.

Fig. 1. Predicted vs Real mean temperatures graphics for LR(top) and RF

Fig.1. illustrates the predicted temperature and the real recorded temperature for year 2015. For the prediction, the output was compared for 4 main cities: Bucharest, Iasi, Cluj-Napoca and Sibiu.

For Iasi, the Linear Regression model demonstrated clear superiority across the available common metrics for the mean temperature (TMED). It achieved a strong fit ($R^2 = 0.8191$), while the complex model's coefficient of determination dropped significantly (0.7116). The error metrics were much lower for the linear approach (MAE 3.2252, RMSE 3.9957) compared to the non-linear one (MAE 4.0831, RMSE 5.0454).

In Cluj-Napoca both models produced very similar results, with LR achieving an R^2 about 0.5% higher for mean temperature (TMED). The error metrics were similarly tight (MAE 2.7138 vs 2.7939).

The performance trend continued in Sibiu, where Linear Regression proved more robust. For TMED, it recorded an R^2 of 0.8026 and an RMSE of 3.9643, significantly better than the RF model's R^2 of 0.7561 and RMSE of 4.4076.

In Bucharest, the complex (Random Forest) model performed marginally better. For TMED, it achieved the highest fit across all cities ($R^2 = 0.8730$, RMSE 3.2033), slightly surpassing the simpler (LR) approach ($R^2 = 0.8530$, RMSE 3.4462). The non-linear method also showed a slight advantage for TMAX and TMIN, although the differences in error metrics were minimal.

CONCLUSIONS

The study found that Linear Regression performed better than Random Forest despite being the simpler model. The Random Forest model tended to overfit the small dataset, learning noise instead of real patterns. Because the data included only temperature values and lacked other important factors like humidity or pressure, the simpler model was more stable and generalized better to new data. This shows that a more complex model is not always better, especially when the dataset is small and limited.