УДК 620.91:004.89

Enhancing predictive accuracy in environmental data analysis: a hybrid LASSO-RFR approach for climatic analysis in Siberia

D.A. Akpuluma, J.I. Abam, C.A. Williams
Scientific Supervisor: Prof., A.V. Yurchenko
Tomsk Polytechnic University, Russia, Tomsk, Lenin str., 30, 634050
E-mail: aa06@tpu.ru, akpoebi@gmail.com

Abstract. *This study introduces a hybrid LASSO-RFR approach for photovoltaic energy forecasting, leveraging LASSO's feature selection with RFR's analytical strength to tackle weather-induced variability. It showcases improved forecast accuracy through simplified datasets and enhanced correlation analysis, resulting in superior model performance. With an MSE of 0.0060 and an R-squared of 85.7% for Model 2, the approach outperforms LASSO-only models, marking a significant advancement in renewable energy analytics and offering a potent forecasting tool for areas with extreme weather.*

Key words*: hybrid machine learning models, advanced statistical modelling, climate data analysis.*

Introduction

Amidst global population growth, the shift from traditional fossil fuels to clean energy is urgent, with photovoltaic (PV) power emerging as a key renewable source. [1] Yet, PV power's dependency on variable weather conditions like sunlight and wind speed necessitates advanced forecasting [2]. Forecasting methods range from statistical and time-series models, which consider weather and historical data, to more complex AI and hybrid models [3,4,6]. A hybrid approach, combining the LASSO and Random Forest Regression (RFR), shows promise for hour-ahead PV power predictions, especially in extreme climates [5]. RFR's strength lies in its ability to process non-linear patterns and large data sets, leading to improved efficiency and reliability in renewable energy forecasting when considering variables such as cloud cover and temperature [6].

This study aims to reveal Siberia's environmental dynamics using a LASSO-Random Forest hybrid model and enhance environmental data analysis methodologies. By integrating LASSO and RFR, it seeks to improve prediction accuracy and refine analytical techniques for complex datasets in extreme weather conditions.

Research methods

In the pursuit of enhancing solar power prediction, various methodologies can be employed, each with its unique strengths and weaknesses.

General methods

Statistical methods like ARIMA are robust for linear trends in historical data but struggle with non-linearities. Machine learning algorithms such as Random Forest and SVM detect non-linearities but can overfit without quality data. Artificial Neural Networks, including CNNs and RNNs, excel in complex pattern modelling but need substantial data and compute power. Physical models offer principled predictions but can oversimplify complex interactions. Hybrid models blend these approaches for improved accuracy but at the cost of increased complexity and computational resources [7].

LASSO-RFR hybrid framework

We present a hybrid LASSO-RFR model for enhanced environmental predictions. LASSO simplifies features, and RFR addresses non-linearity, offering superior forecasts and mitigating overfitting. Data from Tomsk's TOR station, covering solar radiation, temperature, humidity, and wind speed from January 1, 2021–January 23, 2024, served as the basis for our analysis [8].

Equation (1) shows the LASSO regression formula, which combines the OLS loss function with a penalty on the absolute values of coefficients, given as $|a_i|$, regulated by λ. Increasing λ enables LASSO to shrink less important feature coefficients to zero, performing feature selection, enhancing simplicity, and preventing overfitting, thus maintaining interpretability in various climates [7].

$$\text{Loss function} = \text{OLS loss function} + \lambda * \sum_{i=1}^{n} |a_i| . \quad (1)$$

The RFR algorithm, combining decision trees via an ensemble approach with bagging and the random subspace method, excels in classification, regression, and clustering tasks. It constructs multiple trees, each weak alone but collectively strong, to enhance machine learning performance [6].

Results

This paper's findings, derived from applying LASSO-RFR to three models, explore the impact of varying weather parameters on forecast sensitivity in extreme climates. Model 1 incorporates all weather conditions, Model 2 excludes temperature, and Model 3 omits humidity, highlighting the hybrid method's responsiveness to different climatic elements.

Model 1

In this model LASSO-RFR method is used to analyze Siberia's climatic features. LASSO, as a feature selection mechanism, zeroes out less significant parameters, identifying temperature, humidity, wind speed, and month as key predictors (Fig. 1). An RFR model is then trained on these selected features to accurately forecast.

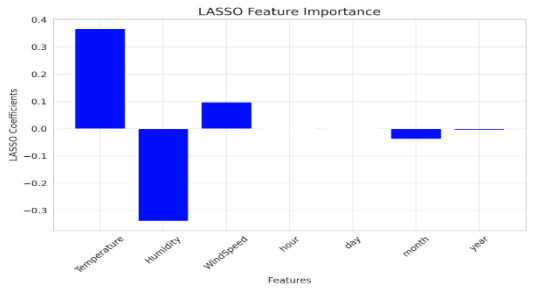

Fig. 1. Plot of Model 1 (normalized data)

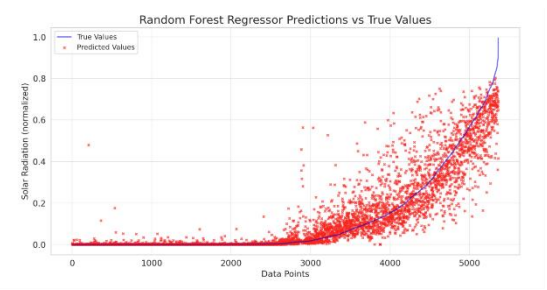

Fig. 2. RFR predictions vs true values (Model 1)

Fig. 2 illustrates RFR's accurate solar radiation predictions (red dots) against actual values (blue line), with an R-squared of 84.6 % and an MSE of 0.0065, showcasing superior performance to LASSO. The model's limitation in capturing peak solar radiation is noted by deviations at peaks.

Model 2

Here, one of the weather parameters is excluded from the analysis, which is temperature. Fig.3 is the chart showing the coefficient values of the features without temperature leaving humidity as the most important feature, while day is the least important feature. In this model there are no zero coefficient features.

Fig. 4 displays true solar radiation (blue line) against predicted values (red dots), showing variance in predictions. The model achieves an R-squared of 85.7 %, explaining most variance, with a reduced MSE of 0.0060, surpassing Model 1's performance.

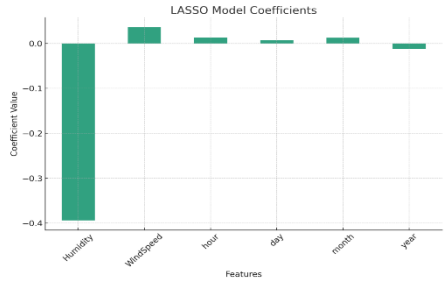

Fig. 3. LASSO feature importance plot of Model 2

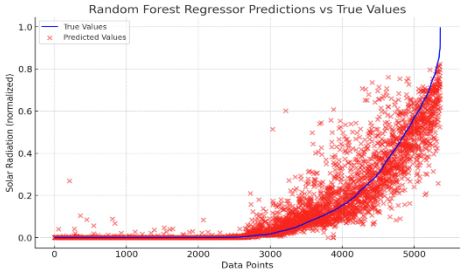

Fig. 4. RFR predictions vs true values (Model 2)

Model 3

In Model 3, excluding humidity, LASSO selected features for RFR analysis. This model yielded an R-squared score of 83.9 % and an MSE of 0.0068, differing slightly from prior models.

Conclusion

The study demonstrates the superior accuracy of a LASSO-RFR hybrid model over LASSO alone for predicting solar radiation in Siberia, significantly enhancing RFR's metrics through LASSO-based feature selection. Notably, Model 2, which excludes temperature from LASSO's input, outperforms other models with an R-squared of 0.857 (85.7 %) and a mean squared error of 0.0060, affirming its efficiency in capturing solar radiation variability. Future directions include refining these models for better performance and interpretability in similar climates. Integrating Explainable Artificial Intelligence (XAI) could further clarify the predictive process, marrying high accuracy with transparency in machine learning models' power output predictions [9].

References

1. Sher F., Curnick O., Azizan M.T. Sustainable conversion of renewable energy sources // Sustainability. – 2021. – Vol. 13, № 5. URL: https://doi.org/10.3390/SU13052940.

2. Kumar D.S., Yagil G.M., Kashyap M., Srinivasan D. Solar irradiance resource and forecasting: a comprehensive review // IET Renewable Power Generation. – 2020. – Vol. 14, № 10 – P. 1641–1656. URL: https://doi.org/10.1049/iet-rpg.2019.1227.

3. Akhter M.N., Mekhilef S., Mokhlis H., Shah N.M. Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques // IET Renewable Power Generation. – 2019. – Vol. 13, № 7 – P. 1009–1023. URL: https://doi.org/10.1049/IET-RPG.2018.5649.

4. Sobri S., Koohi-Kamali S., Rahim N. Abd. Solar photovoltaic generation forecasting methods: A review // Energy Conversion and Management. – 2018. – Vol. 156 – P. 459–497. URL: https://doi.org/10.1016/J.ENCONMAN.2017.11.019.

5. Hengi T., Nussbaum M., Wright M.N., Heuvelink G.B.M., Gräler B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables // PeerJ. – 2018. URL: https://doi.org/10.7717/peerj.5518.

6. Khalyasmaa A., Eroshenko S.A., Chakravarthy T.P., Gasi, V.G., Bollu S.K.Y., Caire R., Atluri S.K.R., Karrolla S. Prediction of solar power generation based on random forest regressor model // International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). – 2019. URL: https://doi.org/10.1109/SIBIRCON48586.2019.8958063.

7. Akpuluma D.A., Yurchenko A.V. Advancing Solar Irradiation Prediction in Extreme Climates: A LASSO Regression Analysis in Tomsk // Proceedings of International Conference on Applied Innovations in IT. – 2024. – Vol. 12, № 1 – P. 257–263. URL: (DOI:Under Indexing).

8. Belan B.D., Ivlev G.A., Sklyadneva T.K. Long-term monitoring of total and UV-B radiation in Tomsk // Atmospheric and Oceanic Optics. – 2012. – Vol. 25, № 4 – P. 281–285.

9. Gunning D., Aha D. DARPA’s Explainable Artificial Intelligence (XAI) Program // AI Magazine. – 2019. – Vol. 40, № 2. URL: https://doi.org/10.1609/aimag.v40i2.2850.