Environmental Engineering and Management Journal xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
http://www.eemj.icpm.tuiasi.ro/; http://www.eemj.eu

"Gheorghe Asachi" Technical University of Iasi, Romania

COMPARITIVE ANALYSIS OF LASSO AND LIGHTGBM MODELS IN PREDICTING PM2.5 CONCENTRATION: BASED ON PROVINCIAL PANEL DATA FROM MAINLAND CHINA, 2011-2020

Xin Ji[1], Wei Xu[1], Yan Yan[1], Rabia Aslam[2], Yanjuan Cui[3]*

[1]School of Management, Shenyang University of Technology, Shenyang, 110870, P.R. China
[2]Department of Commerce, University of the Punjab, Gujranwala, 52250, Pakistan
[3]School of Finance, Dongbei University of Finance and Economics, Dalian, 116025, P.R. China

Abstract

The study conducted a comprehensive analysis of the PM2.5 concentrations and related factors in 31 provinces of mainland China from 2011 to 2020. Through comparing LASSO regression and LightGBM models, this study revealed an important trade-off between model interpretability and prediction accuracy: while LightGBM showed superior predictive performance, LASSO regression provided better interpretability of factor relationships. The results showed that, in terms of overall predictive performance, the LightGBM model demonstrated higher accuracy and stronger generalization capabilities in predicting PM2.5 concentrations. This superiority was reflected in key performance indicators compared to the LASSO model. However, the LightGBM model's "black box" nature limited its ability to explain the mechanisms behind PM2.5 variations, whereas the LASSO model offered clear insights into factor relationships despite lower overall accuracy. By optimizing the LASSO model and employing cluster analysis, the study significantly improved the model's predictive capability by accounting for regional environmental similarities. The cluster-optimized LASSO model even outperformed the LightGBM model in predicting the PM2.5 concentrations in certain regions, demonstrating the effectiveness of combining statistical learning with domain knowledge. Nevertheless, the LASSO model, supported by data from a single province, was prone to overfitting due to the limited scale of training data and environmental heterogeneity, resulting in deviation from the true trend of PM2.5 concentrations. The study not only proposed an accurate method for predicting annual average PM2.5 concentrations but also provided insights into the balance between model accuracy and interpretability, offering policymakers both precise predictions and clear factor relationships for informed decision-making.

Key words: LASSO regression, LightGBM, panel data, PM2.5 prediction

Received: April, 2024; Revised final: January, 2024; Accepted: January, 2025

* Author to whom all correspondence should be addressed: e-mail: 1526343937@qq.com; Phone: +86 0411-84710221; Fax: +86 0411-84712503