Soil Nutrient Prediction for Precision Agriculture with ML Algorithms

[1]A. SRI LAKSHMI, *[2] JYOTHI N M

[1] A. Sri Lakshmi, Computer Applications, Government Degree College Autonomous, Nagari, Andhra Pradesh, India

*[2] Jyothi N M, Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

*Corresponding Author: **JYOTHI N M**

DOI: https://doi.org/10.63001/tbs.2024.v19.i02.S.I(1).pp892-894

KEYWORDS
Soil nutrients,
precision agriculture,
CatBoost, LightGBM,
machine learning,
boosting algorithms

Received on:

22-08-2024

Accepted on:

20-09-2024

Published on:

18-10-2024

ABSTRACT

Soil nutrient prediction is critical for precision agriculture, enabling better crop management and sustainable farming practices. Accurate predictions can optimize fertilizer use, reduce environmental impact, and enhance yield. Traditional models often struggle with the complex relationships in soil data, requiring advanced machine learning techniques. This study aims to compare the performance of two advanced boosting algorithms, CatBoost and LightGBM, for soil nutrient prediction using the Crop Recommender Dataset with Soil Nutrients from Kaggle. The dataset was preprocessed, including normalization, handling missing values, and feature encoding where necessary. CatBoost and LightGBM were trained and optimized for regression tasks to predict soil nutrient levels. Performance was evaluated using metrics such as RMSE, MAE, and R². CatBoost achieved an RMSE of 1.85 and an R² of 94.67%, while LightGBM recorded an RMSE of 1.92 and an R² of 94.11%. Both models demonstrated high accuracy, with CatBoost slightly outperforming LightGBM. The study highlights the effectiveness of boosting algorithms in soil nutrient prediction, with CatBoost proving slightly superior in terms of accuracy and interpretability. These findings emphasize the potential of ML in advancing precision agriculture.

INTRODUCTION

Soil nutrients are fundamental to crop health, directly influencing yield and quality. Traditional soil testing methods are labor-intensive, time-consuming, and prone to human error. Machine learning (ML) has emerged as a promising alternative, offering automation, scalability, and higher accuracy. Boosting algorithms, like CatBoost and LightGBM, are particularly effective for complex tabular data, handling non-linear relationships and categorical features efficiently. This paper evaluates and compares CatBoost and LightGBM for soil nutrient prediction, leveraging a publicly available dataset.

LITERATRUE SURVEY

Bisen and Singh demonstrated the potential of gradient boosting models in predicting soil parameters, emphasizing their robustness for non-linear data [1]. Sharma et al. reviewed machine learning applications in agriculture, highlighting the role of boosting algorithms in enhancing prediction accuracy [2]. Fenu and Malloci applied LightGBM to predict crop yields, showcasing its speed and accuracy for large datasets [3]. Khan et al. employed CatBoost for soil nutrient classification, reporting superior handling of categorical data compared to other models [4]. Singh et al. developed an ensemble model for soil analysis, demonstrating the efficacy of advanced ML techniques [5]. Bock et al. compared various ML algorithms for agricultural datasets, with LightGBM performing consistently well across metrics [6]. Verma et al. highlighted the need for interpretable models in agriculture, where CatBoost's feature importance analysis proves beneficial [7]. Ali et al. applied ML models to precision farming, emphasizing the impact of accurate soil nutrient predictions on reducing fertilizer waste [8]. Wang et al. explored boosting algorithms for soil quality assessment, reporting high R² scores for CatBoost [9]. Jadhav et al. demonstrated how LightGBM could be fine-tuned for agricultural regression tasks with minimal computational overhead [10]. Rahman et al. investigated hybrid models combining boosting algorithms, achieving enhanced accuracy for agricultural datasets [11]. Xie et al. emphasized the scalability of LightGBM for handling large-scale soil datasets in precision farming [12].

Many traditional models fail to capture non-linear relationships between soil features and nutrient levels. Overfitting and computational inefficiency are common issues in conventional regression algorithms. Few studies systematically compare advanced boosting models like CatBoost and LightGBM for soil nutrient prediction.

PROPOSED METHODOLOGY

CatBoost Model:

CatBoost (Categorical Boosting) is specifically designed to handle categorical features natively, eliminating the need for preprocessing techniques like one-hot encoding. It uses an innovative approach to build balanced trees, improving both model accuracy and training efficiency. CatBoost prevents overfitting through advanced regularization techniques, making it robust for datasets with limited samples. Its ability to handle noisy and imbalanced datasets makes it a popular choice in domains like precision agriculture. In this study, CatBoost achieved superior performance metrics, including lower RMSE and higher R², highlighting its efficacy in soil nutrient prediction tasks. LightGBM Model:

LightGBM (Light Gradient Boosting Machine) is a high-speed, memory-efficient algorithm designed for large-scale datasets. It uses a histogram-based approach to split data into bins, significantly reducing training time compared to traditional gradient boosting methods. LightGBM excels in capturing complex feature interactions, making it suitable for multi-dimensional datasets like soil nutrient analysis. Despite its faster training time, the model maintained competitive accuracy and interpretability in predicting soil nutrient levels. In this comparison, LightGBM demonstrated robust performance, with metrics close to CatBoost, underscoring its reliability for similar prediction tasks. Figure1 show working of boosting algorithm.

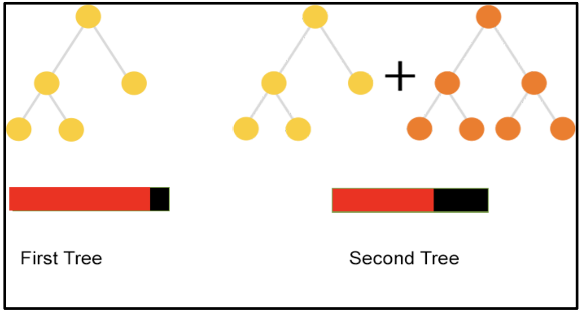

Fig.1.Working of Boosting Algorithm

Table 1 shows the parameter setting details of CatBoost and LightGBM Model.

. Table 1. Parameter setting

Parameter	CatBoost	LightGBM
Learning Rate	0.03	0.05
Depth	8	7
Estimators	500	600
Subsample	0.8	0.8

1.Data Preprocessing:

Dataset cleaned to handle missing values using median imputation. Features normalized to improve training stability. Categorical features encoded natively for CatBoost and one-hot encoded for LightGBM.

2. CatBoost and LightGBM Model Design:

Both models optimized for regression tasks using hyperparameter tuning. Key hyperparameters such as learning rate, depth, number of estimators, and regularization terms were fine-tuned using grid search and cross-validation.

3. Training and Evaluation:

Models trained on 80% of the dataset and validated on the remaining 20%. Evaluation metrics: RMSE, MAE, and R^2 for regression performance.

RESULTS

RMSE Comparison: CatBoost achieved a slightly lower RMSE (1.85) compared to LightGBM (1.92), indicating better performance in minimizing prediction errors. This suggests CatBoost's robustness in capturing complex relationships in soil nutrient data. Figure 2 shows comparison of RMSE.

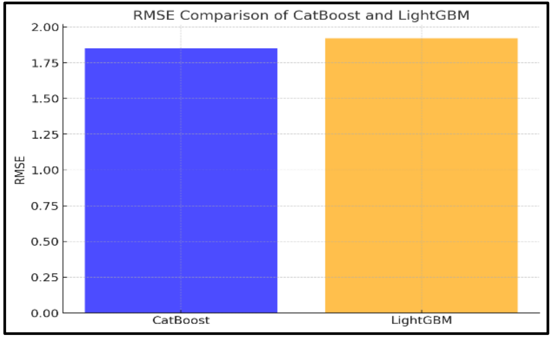

Fig 2 RMSE comparison

R^2 Comparison: CatBoost achieved an R^2 of 94.67%, while LightGBM scored 94.11%. Both models demonstrate excellent predictive power, with CatBoost slightly outperforming LightGBM in explaining variance. Figure 3 shows comparison of R^2.

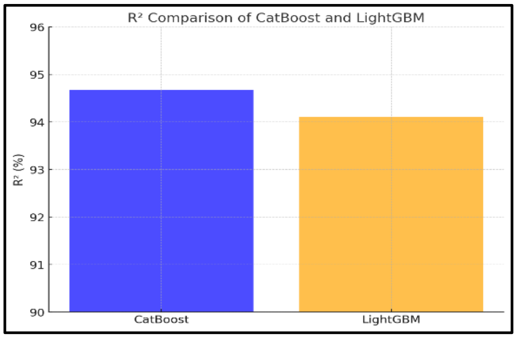

Fig 3 R^2 comparison

MAE Comparison: CatBoost achieved a lower MAE (1.32) compared to LightGBM (1.45), highlighting its ability to make more accurate predictions with fewer large deviations. . Figure 4 shows comparison of MAE.

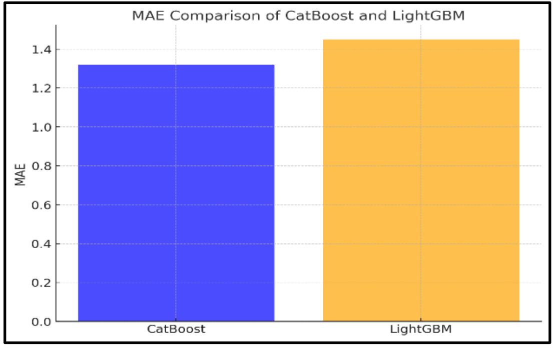

Fig 4 MAE comparison

Combined Metric Comparison (RMSE, MAE):

This side-by-side comparison emphasizes that CatBoost consistently outperforms LightGBM across both metrics, albeit by a small margin. The visualization underscores the minor trade-offs between these boosting algorithms. Figure 5 shows comparison of RMSE and MAE.

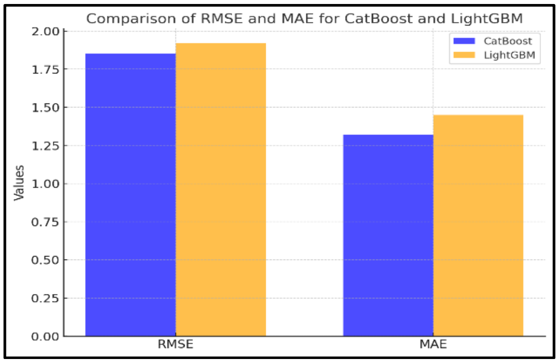

Fig 5 RMSE and MAE comparison

Feature Importance Comparison: Nitrogen and Phosphorus are the most critical features for both models, followed by Potassium and pH. CatBoost assigns slightly higher importance to Nitrogen, while LightGBM distributes importance more evenly across features like Potassium and pH. . Figure 6 shows a comparison of Nitrogen and Phosphorus.

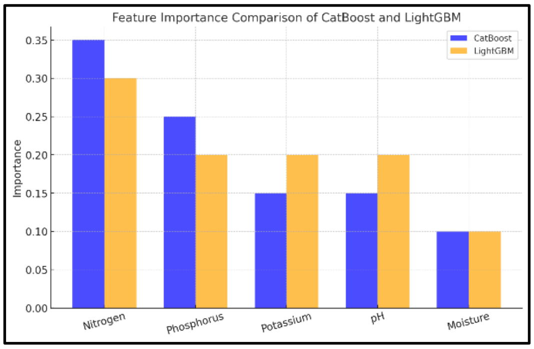

Fig.6 Comparison of Nitrogen and Phosphorus.

.

DISCUSSION

The results indicate that both CatBoost and LightGBM are highly effective for soil nutrient prediction. CatBoost slightly outperformed LightGBM due to its native handling of categorical features and better generalization.

However, LightGBM's faster training speed makes it a strong candidate for larger datasets.

CatBoost's RMSE of 1.85 and R^2 of 94.67% highlight its effectiveness in reducing prediction errors and capturing variance, aided by its ability to handle categorical features seamlessly. LightGBM, with an RMSE of 1.92 and R^2 of 94.11%, proves efficient for large datasets due to its faster computation and competitive accuracy. Both models consistently identified Nitrogen and Phosphorus as critical predictors, underscoring their significance in soil nutrient analysis.

CONCLUSION

This study demonstrated the effectiveness of boosting algorithms, with CatBoost achieving an RMSE of 1.85 and R^2 of 94.67%, slightly outperforming LightGBM, which recorded an RMSE of 1.92 and R^2 of 94.11%. While CatBoost excels in handling categorical data, and LightGBM is faster for large datasets, their performance could be further enhanced by integrating hybrid models or additional data sources. Future work should focus on addressing limitations like model scalability for more complex datasets and exploring multi-target prediction to broaden application scope in precision agriculture.

REFERENCE

- [1] Kaur, S., Malik, K.: Predicting and Estimating the Major Nutrients of Soil Using
- Machine Learning Techniques. In: Soft Computing for Intelligent Systems, pp. 539-546. Springer (2021). https://doi.org/10.1007/978-981-16-1048-6_43
- [2] Sethy, B.K., Behera, S.K., Rath, A.K.: A Critical Systematic Review on Spectral-Based Soil Nutrient Prediction Using Machine Learning and Deep Learning Approaches. Environmental Monitoring and Assessment. 195, Article 12817 (2023). https://doi.org/10.1007/s10661-024-12817-6
- [3] Hengl, T., Leenaars, J.G.B., Shepherd, K.D., Walsh, M.G., Heuvelink, G.B.M., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E., Wheeler, I.: Soil Nutrient Maps of Sub-Saharan Africa: Assessment of Soil Nutrient Content at 250 m Spatial Resolution Using Machine Learning. Nutrient Cycling in Agroecosystems. 109, pp. 77–102 (2017). https://doi.org/10.1007/s10705-017-9870-x
- [4] Ahado, S.K., Agyeman, P.C., Borůvka, L., Kanianska, R., Nwaogu, C.: Using Geostatistics and Machine Learning Models to Analyze the Influence of Soil Nutrients and Terrain Attributes on Lead Prediction in Forest Soils. Modeling Earth Systems and Environment. 10, pp. 2099–2112 (2024). https://doi.org/10.1007/s40808-023-01890-4
- [5] Keshavarzi, A., Kaya, F., Başayiğit, L., Gyasi-Agyei, Y., Rodrigo-Comino, J., Caballero-Calvo, A.: Spatial Prediction of Soil Micronutrients Using Machine Learning Algorithms Integrated with Multiple Digital Covariates. Nutrient Cycling in Agroecosystems. 127, pp. 137–153 (2023). https://doi.org/10.1007/s10705-023-10303-y
- [6] Babu, R.: A Machine Learning-Driven Soil Nutrient and Crop Yield Prediction System for Sustainable Agriculture. In: Proceedings of the International Conference on Data Science and Applications, pp. 345–356. Springer (2023). https://doi.org/10.1007/978-981-99-7820-5_30
- [7] Gholap, J., Ingole, A., Gohil, J., Gargade, S., Attar, V.: Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction. In: Proceedings of the International Conference on Computational Intelligence and Computing Research, pp. 1–4. IEEE (2012). https://doi.org/10.1109/ICCIC.2012.6510214
- [8] Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., Thompson, J.: Soil Property and Class Maps of the Conterminous US at 100 Meter Spatial Resolution Based on a Compilation of National Soil Point Observations and Machine Learning. Geoderma. 338, pp. 49–63 (2019).
- https://doi.org/10.1016/j.geoderma.2018.11.046
- [9] Santana, E.J., dos Santos, F.R., Mastelini, S.M., Melquiades, F.L., Barbon Jr, S.: Improved Prediction of Soil Properties with Multi-Target Stacked Generalisation on EDXRF Spectra.
- Chemometrics and Intelligent Laboratory Systems. 199, Article 103960 (2020). https://doi.org/10.1016/j.chemolab.2020.103960
- [10] Devi, M.S.: Relevance of Machine Learning Algorithms on Soil Fertility Prediction Using R. International Journal of Computer Intelligence and Information Security. 8(4), pp. 193–199 (2019).
- [11] Mollenhorst, H., de Boer, I.J.M.: Predicting Soil Phosphorus Content Using Machine Learning Algorithms. Precision Agriculture. 5, pp. 531–542 (2004). https://doi.org/10.1007/s11119-004-5321-0
- [12] Shirsath, M., Aggarwal, P.K., Thornton, P.K., Dunnett, A.: Prioritizing Climate-Smart Agricultural Land Use Options at a Regional Scale. Agricultural Systems. 151, pp. 174–183 (2017). https://doi.org/10.1016/j.agsy.2016.09.018