Machine Learning-Based Prediction of Carbon Dioxide Emissions: A Comparative Analysis of Ensemble Models and Feature Importance Evaluation

[1]Ruimin Ma, [2]Qiong Li, [3,*]Alexander Kovshov

[1,2,3]St. Petersburg State University, Russia,
Email: [1]mrm1999026@gmail.com, [2]lq17829111293@outlook.com , [3]a.kovshov@spbu.ru
*Corresponding Author: Alexander Kovshov, (Email:a.kovshov@spbu.ru)

Abstract- Accurate forecasting of carbon dioxide (CO_2) emissions is crucial for developing effective environmental policies and mitigating climate change. In this study, we apply machine learning models, including Random Forest, XGBoost, LightGBM, and CatBoost, to predict CO_2 emissions based on a dataset covering 107 countries from 2000 to 2020. We investigate the influence of key economic, social, environmental, and energy-related factors on CO_2 emissions and assess the predictive performance of each model. To enhance interpretability, we employ feature importance analysis to identify the most significant drivers of CO_2 emissions. By leveraging Permutation Importance, we quantify the contribution of various features across different models. Our methodology integrates a time-window-based feature engineering approach, allowing us to capture temporal patterns in CO_2 emissions trends. Experimental results show that CatBoost delivers the highest overall predictive performance, benefiting from its Ordered Boosting mechanism and superior handling of categorical data. LightGBM and XGBoost also achieve strong results, with XGBoost demonstrating notable advantages in controlling prediction bias. The feature importance analysis highlights the dominant role of energy-related factors, particularly electricity consumption from fossil fuels and renewables, in shaping CO_2 emissions. Additionally, social and economic indicators, such as land area and GDP growth, exhibit varying levels of impact across models. This study underscores the efficacy of machine learning techniques in CO_2 emissions forecasting and provides valuable insights into the underlying drivers of emissions. The findings contribute to advancing data-driven environmental policy-making.

Keywords: Carbon Dioxide Emissions Prediction, Machine Learning, Ensemble Learning Models, Permutation Importance.

I. INTRODUCTION

Rising atmospheric CO_2 concentration is a primary driver of global climate change and threatens ecosystems, economic activity, and social welfare. Although climate change manifests as an environmental problem, its causes are closely linked to heterogeneous development paths and energy structures across countries. In the context of the Paris Agreement and carbon-neutrality commitments, reliable CO_2 emissions prediction is crucial for designing mitigation strategies and evaluating policy scenarios.

Traditional statistical approaches, such as multiple regression, often struggle with nonlinear relationships, high-dimensional feature spaces, and complex interactions between economic growth, energy use, and demographic factors. Machine learning (ML), and in particular tree-based ensemble models, can capture such nonlinearities and interactions while offering competitive predictive performance.

Despite a growing body of ML-based environmental studies, there remains a shortage of systematic comparisons of ensemble algorithms using a unified multi-country dataset, combined with a transparent assessment of feature importance. To address this gap, we: (1) Construct a joint dataset for 107 countries over 2000–2020, retaining 14 key factors with established relevance to CO_2 emissions. (2) Implement a 5-year sliding-window feature engineering scheme to incorporate short-term temporal dynamics. (3) Compare four ensemble models-Random Forest, XGBoost, LightGBM, and CatBoost-in terms of multiple error metrics. (4) Quantify feature contributions via permutation importance, aggregated across time lags.

The study aims to provide both a robust predictive benchmark and interpretable evidence on the main drivers of CO_2 emissions at the global scale.

II. METHODOLOGY

2.1. Ensemble learning models

We consider four tree-based ensemble models that are widely used in tabular regression tasks and known for robust performance and interpretability.

2.1.1. Random Forest (RF): Random Forest constructs an ensemble of decision trees trained on bootstrap samples of the data. At each split, only a random subset of features is considered, which decorrelates trees and improves generalization. The final prediction is the average of all tree outputs. RF is simple, robust to noise, and relatively insensitive to hyperparameters.

2.1.2. XGBoost:

XGBoost is a gradient boosting framework that builds trees sequentially, with each new tree fitted to the residuals of the current ensemble. It uses second-order information (gradients and Hessians) to optimize a regularized objective, controlling both training loss and model complexity. Learning rate and regularization parameters help prevent overfitting and improve stability.

2.1.3. LightGBM:

LightGBM is a gradient boosting method optimized for efficiency on large, high-dimensional datasets. It employs histogram-based feature binning and leaf-wise tree growth with depth constraints, significantly reducing training time and memory usage. LightGBM also supports advanced regularization and handles sparse features effectively.

2.1.4. CatBoost:

CatBoost is a gradient boosting method with specific optimizations for categorical features and small datasets. It uses ordered boosting to reduce target leakage, symmetric tree structures for efficient implementation, and specialized encodings for categorical variables. These design choices improve generalization and reduce overfitting, making CatBoost particularly effective on heterogeneous tabular data.

2.2. Permutation importance and temporal aggregation

To interpret the trained models, we adopt permutation importance. For a given trained model and evaluation set, the importance of feature x_i is defined as the increase in prediction error when the values of x_i are randomly permuted while all other features are kept fixed. A larger performance degradation implies greater importance of that feature.

Because our input uses a 5-year sliding window, each original variable x_i appears at time lags t, t-1, ...,t-4. We aggregate its importance over the window by summing the importance scores across lags:

$$\text{Importance}(x_i) = \sum_{k\in\{0,1,2,3,4\}} \text{Importance}\,(x_{i,t-k}),$$

which yields a single contribution score per conceptual feature. This provides an interpretable ranking of long-short-term influences on CO_2emissions.

III. EXPERIMENTAL SETUP

3.1. Dataset and preprocessing:

We use a dataset covering 107 countries from 2000 to 2020, resulting in 2247 samples after preprocessing. The variables include 14 factors capturing economic, demographic, geographic, energy, and technological characteristics, such as GDP growth, GDP per capita, population, land area, latitude, access to electricity, electricity generation from fossil fuels and renewables, low-carbon electricity share, primary energy consumption per capita, energy intensity, and renewable electricity capacity per capita.

Data preprocessing proceeds as follows:

1. Data cleaning: missing values and obvious outliers are treated using statistical techniques and domain knowledge to improve consistency.
2. Feature selection: based on prior literature and preliminary analysis, we retain 14 core predictors and discard variables with weak or redundant contributions.
3. Normalization: we apply Min-Max scaling to mitigate scale differences and stabilize model training.
4. Joint modeling: country-level time series are pooled into a single dataset to train one unified global model.
5. Data splitting: the data are partitioned into training, validation, and test sets in a 6:2:2 ratio to enable unbiased performance assessment.

3.2. Prediction target and time-window design

To predict annual CO_2 emissions, we employ a 5-year sliding window. For each target yeart, the input vector concatenates the values of the 14 features for years t−4 to t (inclusive), and the output is the CO_2emissions at year t.

This window length balances two considerations:

1. it is long enough to capture short-term dynamics and structural changes in energy use and economic activity.
2. it remains compact, limiting dimensionality and reducing overfitting risk. The choice is also consistent with common practices in time series forecasting where medium-length windows are used to capture recent trends.

3.3. Evaluation metrics

We evaluate models on the test set using five standard metrics for regression: coefficient of determination R^2, mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE).Together, these metrics provide a comprehensive view of global fit, error magnitude, and relative error behavior.

IV. RESULTS AND DISCUSSION

4.1. Learning curves

We illustrate learning dynamics using the CatBoost model as a representative example (Figures 1 and 2). During training, both training and validation R^2 increase and stabilize around 0.93, while RMSE decreases and converges to approximately 2.73×10^5. The close alignment of training and validation curves suggests good generalization and limited overfitting. Random Forest does not naturally provide a meaningful epoch-based learning curve because trees are grown on bootstrap samples in a single stage rather than via iterative residual fitting.

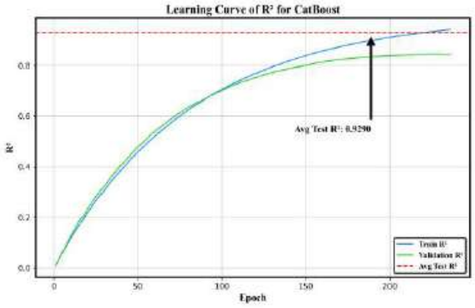

Figure 1: Learning Curve of R^2 for CatBoost

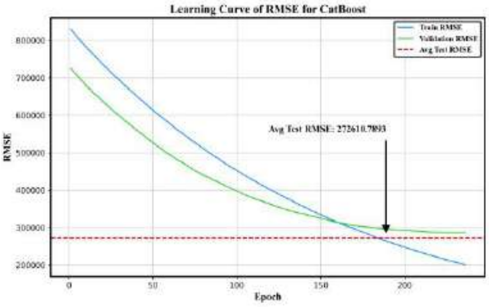

Figure 2: Learning Curve of RMSE for CatBoost

4.2. Comparative prediction performance

Table 1 summarizes the performance of all four models on the test set. The CatBoost model attains the highest R^2 (0.9290) and the lowest MSE/RMSE, indicating superior overall fit and error control. LightGBM performs comparably, with slightly lower R^2 and somewhat higher error metrics. Both models clearly outperform Random Forest and XGBoost in terms of global goodness of fit.

Table 1: Comparison of model performance.

Model	**Avg_R^2**	**Avg_MSE**	**Avg_RMSE**	**Avg_MAE**	**Avg_MAPE**
Random Forest	0.8919	113337578108.78	336656.47	73959.43	2022.23
XGBoost	0.8936	111403327458.13	333771.37	53801.34	132.25
LightGBM	0.9232	80463645220.44	283661.15	57984.45	271.15
CatBoost	0.9290	74316642461.03	272610.79	60572.95	2108.96

Interestingly, XGBoost achieves the best MAE and MAPE, implying that it handles local prediction errors and relative deviations particularly well, which may be advantageous in applications where bias control at the individual-country level is critical. Random Forest yields the weakest performance across most metrics but still reaches an R^2close to 0.89, making it a reasonable baseline and a robust, easy-to-tune model for preliminary analysis.

Overall, CatBoost and LightGBM appear to be strong default choices for CO_2 emissions forecasting with this type of data, while XGBoost can be preferred when minimizing local bias is the primary objective.

4.3. Feature importance analysis

Permutation importance scores for the 14 features are computed for all four models. Table 2 reports the aggregated importance values, highlighting the most influential variables.

Table 2. Feature importance values for different models

Feature	**Random Forest**	**XGBoost**	**LightGBM**	**CatBoost**
GDP growth	2.5679e-04	-4.4279e-03	-4.5251e-06	3.7521e-02
GDP per capita	0.0000e+00	1.0013e-02	7.4584e-04	2.0622e-02
Population density	0.0000e+00	7.7472e-04	8.5298e-05	3.5378e-02
Land area	3.2897e-02	1.2336e+00	1.3272e-01	2.9307e-01
Latitude	1.6045e-01	9.6192e-06	1.4167e-01	3.1739e-02
Renewable energy share	-5.5275e-05	2.5860e-06	3.4001e-04	3.0777e-02
Access to electricity	0.0000e+00	2.3961e-05	1.3213e-05	4.1829e-02
Access to clean fuels for cooking	0.0000e+00	5.8716e-06	1.7262e-04	2.8613e-02
Electricity from fossil fuels	3.5012e-01	8.2482e-02	7.8023e-01	4.0250e-01
Electricity from renewables	1.1424e-01	1.3351e-04	3.4691e-01	1.2907e-01
Low carbon electricity	0.0000e+00	5.4632e-03	1.1704e-05	2.7701e-02
Primary energy consumption per capita	8.7420e-05	0.0000e+00	1.2355e-02	5.4914e-02
Energy intensity level of primary energy	1.1105e-05	1.2337e-04	1.0783e-04	2.8128e-02
Renewable electricity generating capacity per capita	0.0000e+00	0.0000e+00	4.9166e-04	4.0469e-02

"Electricity from fossil fuels" has the highest importance in all models, especially in LightGBM (≈ 0.78) and CatBoost (≈ 0.40). This confirms its dominant role in explaining CO_2 emissions. "Electricity from renewables" and "Low-carbon electricity" also exhibit substantial contributions, though generally lower than fossil-fuel electricity.

"Land area" shows consistently high importance, particularly in XGBoost, where it receives the largest score among all features. "Latitude" is also influential in Random Forest and LightGBM, suggesting that geographic position-and associated climate conditions, energy demand, and infrastructure patterns-meaningfully affect emissions trajectories.

GDP-related variables and population measures have more moderate yet non-negligible importance. In CatBoost, "GDP growth" and "GDP per capita" contribute positively to prediction accuracy, whereas XGBoost assigns a slightly negative importance to GDP growth, reflecting a complex, model-dependent relationship between economic expansion and emissions.

Variables such as "Renewable energy share", "Energy intensity of primary energy", and "Renewable electricity generating capacity per capita" generally show smaller but non-zero importance. In some models, renewable-related features have negative permutation scores, indicating that their effects may be intertwined with other predictors or that higher renewable shares coincide with structural changes in the energy system.

The feature importance visualization for CatBoost (Figure 3) shows a clear ranking pattern, with "Electricity from fossil fuels", "Land area", and "Electricity from renewables" as the top three features. This aligns with the view that emissions are primarily shaped by energy structure, physical scale, and spatial characteristics, modulated by socioeconomic factors.

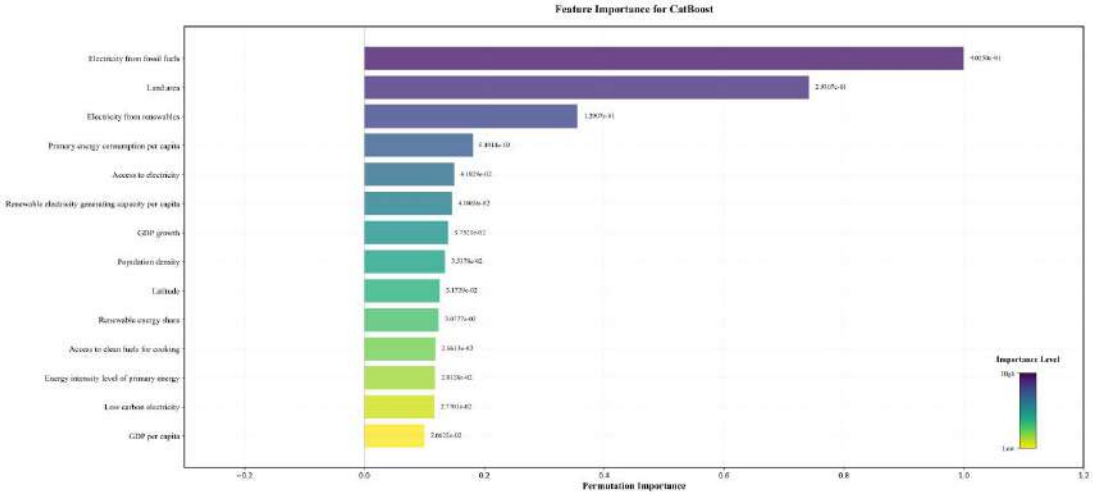

Figure 3: Feature Importance for CatBoost

V. CONCLUSION

This paper investigates the use of ensemble learning models for predicting CO_2 emissions using a multi-country panel dataset spanning 107 countries from 2000 to 2020. By combining a 5-year sliding-window feature representation with four tree-based models (Random Forest, XGBoost, LightGBM, and CatBoost).We systematically evaluate predictive performance and interpret the main drivers of emissions via permutation importance.

CatBoost delivers the best overall performance, with the highest $R^2$and lowest RMSE, followed closely by LightGBM. Both models effectively capture nonlinear interactions in the data and show strong generalization. XGBoost performs competitively and achieves the smallest MAE and MAPE, making it attractive for applications focusing on relative error control. Random Forest is less accurate but remains a robust benchmark.

Energy-related variables, especially electricity generation from fossil fuels, are the most important predictors across all models. Electricity from renewables, land area, and latitude also play key roles, reflecting the combined influence of energy structure, physical scale, and geography on emissions patterns. Economic and social indicators further modulate emissions but exhibit more model-dependent effects.

The results demonstrate that ensemble learning provides reliable and interpretable tools for CO_2 emissions forecasting. The prominence of fossil-fuel electricity underscores the importance of decarbonizing power systems, while the significance of land area and latitude suggests that geographically tailored mitigation strategies are needed.

Future work may extend this framework by incorporating additional explanatory variables, exploring hybrid models that combine machine learning with domain-specific constraints, and performing regional analyses to uncover heterogeneous effects across country groups.

REFERENCES

[1] Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10), 1340-1347.

[2] Amiri-Ramsheh, B., Larestani, A., Atashrouz, S., Nasirzadeh, E., Essakhraoui, M., Abedi, A., ... & Hemmati-Sarapardeh, A. (2025). Toward accurate prediction of carbon dioxide (CO2) compressibility factor using tree-based intelligent schemes (XGBoost and LightGBM) and equations of state. Results in Engineering, 25, 104035.

[3] As, M., & Bilir, T. (2024). Machine learning algorithms for energy efficiency: Mitigating carbon dioxide emissions and optimizing costs in a hospital infrastructure. Energy and Buildings, 318, 114494.

[4] Faruque, M. O., Rabby, M. A. J., Hossain, M. A., Islam, M. R., Rashid, M. M. U., & Muyeen, S. M. (2022). A comparative analysis to forecast carbon dioxide emissions. Energy Reports, 8, 8046-8060.

[5] Gür, T. M. (2022). Carbon dioxide emissions, capture, storage and utilization: Review of materials, processes and technologies. Progress in Energy and Combustion Science, 89, 100965.

[6] Li, S., Siu, Y. W., & Zhao, G. (2021). Driving factors of CO2 emissions: further study based on machine learning. Frontiers in Environmental Science, 9, 721517.

[7] Li, X., & Zhang, X. (2023). A comparative study of statistical and machine learning models on carbon dioxide emissions prediction of China. Environmental Science and Pollution Research, 30(55), 117485-117502.

[8] Mwakipunda, G. C., Ibrahim, A. W., Kouassi, A. K. F., Komba, N. A., Ayimadu, E. T., Mgimba, M. M., ... & Yu, L. (2024). Estimating carbon dioxide solubility in brine using mixed effects random forest based on genetic algorithm: implications for carbon dioxide sequestration in saline aquifers. SPE Journal, 29(11), 6530-6546.

[9] Namboori, S. (2020). Forecasting carbon dioxide emissions in the United States using machine learning (Doctoral dissertation, Dublin, National College of Ireland).

[10] Qin, J., & Gong, N. (2022). The estimation of the carbon dioxide emission and driving factors in China based on machine learning methods. Sustainable Production and Consumption, 33, 218-229.

[11] Tanwar, A. (2023). Global Data on Sustainable Energy (2000-2020). Kaggle.

[12] Wu, C., Ju, Y., Yang, S., Zhang, Z., & Chen, Y. (2023). Reconstructing annual XCO2 at a 1 km× 1 km spatial resolution across China from 2012 to 2019 based on a spatial CatBoost method. Environmental Research, 236, 116866.

[13] WU, Y., DUAN, Q., & SUI, J. (2024). PREDICTION OF CARBON DIOXIDE LEVELS IN THE EUROPEAN ALPS BASED ON MACHINE LEARNING ALGORITHMS. APPLIED AND COMPUTATIONAL ENGINEERING Учредители: EWA Publishing, 95(1), 23-33.