Forecasting Carbon Emissions in Thailand Using Machine Learning

Kaung Htet Cho
Mechatronics and Machine Intelligence
Asian Institute of Technology
Pathumthani, Thailand
st124092@ait.asia

Htoo Min Htet
Mechatronics and Machine Intelligence
Asian Institute of Technology
Pathumthani, Thailand
st125298@ait.asia

Thang Sian Khawm
Information & Communications Technologies
Asian Institute of Technology
Pathumthani, Thailand
st124642@ait.asia

Saw Kapaw Say
Internet of Things (IoT) Systems Engineering
Asian Institute of Technology
Pathumthani, Thailand
st124090@ait.asia

Abstract

This study presents a machine learning-based approach to forecasting carbon dioxide (CO_2) emissions in Thailand, a country striving to achieve carbon neutrality by 2050 despite its current reliance on fossil fuels. Using the OWID-CO2 dataset and additional energy indicators, we trained five models—XGBoost, Random Forest, LightGBM, CatBoost, and Gradient Boosting—on global and ASEAN data (excluding Thailand) and evaluated them solely on Thai emissions data from 1950–2023. Our results show that XGBoost achieved the highest predictive accuracy with an R^2 score of 0.9968, demonstrating strong generalization capability. Feature importance analysis highlighted GDP and environmental indicators such as temperature change from N_2O as key predictors. In addition to informing national climate policy, the findings help identify business opportunities in renewable energy, energy efficiency, and carbon offset sectors, supporting Thailand's transition to a low-carbon economy.

1. Introduction

Climate change is one of the most pressing global challenges, and carbon dioxide (CO_2) emissions from fossil fuel consumption are a leading contributor to global warming. Thailand, like many developing countries, is facing the dual challenge of economic growth and environmental sustainability. In response to global climate goals, Thailand has committed to achieving carbon neutrality by 2050 [1]. However, this goal is complicated by the country's continued reliance on fossil fuels, particularly in its energy and transportation sectors.

This research aims to support Thailand's transition toward a low-carbon economy by leveraging machine learning to forecast future CO_2 emission trends. The study focuses on two key research questions:

1. What are the future CO_2 emission trends in Thailand?
2. What are the business opportunities in the carbon reduction space?

To address these questions, we use the publicly available *OWID-CO2 dataset* from Our World in Data [3], which provides historical CO_2 emissions and energy production data. We preprocess the dataset by integrating additional energy indicators to create feature-rich training inputs. Multiple machine learning algorithms are then evaluated, including Random Forest, LightGBM, Gradient Boosting, CatBoost and XGBoost. Our preliminary results show that XGBoost outperforms other models in predicting CO_2 emission trends in Thailand.

2. Literature Review

Predicting carbon emissions has become an increasingly important area of research, especially with growing global commitments to reduce greenhouse gas (GHG) emissions. Numerous studies have applied machine learning (ML) techniques to estimate and forecast CO_2 emissions at national and global levels.

For instance, Zhao et al. [4] used support vector regression and long short-term memory (LSTM) models to predict China's CO_2 emissions, demonstrating the potential of

time-series-based deep learning for long-range forecasting. Similarly, Al-Ghezi et al. [2] explored multiple ML algorithms to estimate CO_2 emissions based on economic indicators, finding that ensemble models tend to outperform linear models due to their ability to capture non-linear relationships.

Despite these advancements, several critical gaps remain, particularly in the context of developing countries like Thailand:

- Most prior studies focus on large economies or global datasets, with limited research specific to Thailand and Southeast Asia.
- There is limited research evaluating how machine learning models can generalize across countries and predict emissions accurately without access to national historical data.
- Few studies link emission forecasting to actionable insights for policy design or economic opportunities in the sustainability sector.

This study addresses these gaps by developing a machine learning-based CO_2 emissions forecasting model tailored specifically to Thailand. It uses comprehensive historical data and global context to assess national emissions trends and supports business and policy strategies aligned with Thailand's carbon neutrality goals.

3. Methodology

This section presents the methodology used to predict CO_2 emissions in Thailand using machine learning, guided by a clean data science pipeline and a real-world deployment architecture. Our process is divided into two core phases: (1) model development and (2) deployment.

3.1. Overall Workflow

Figure 1 illustrates the main steps in the machine learning workflow, from data acquisition to model training.

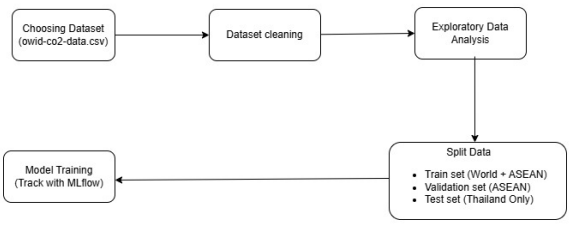


Figure 1. Machine Learning Workflow for CO_2 Emission Prediction

3.2. Dataset and Sources

The dataset used is the OWID-CO2 dataset published by Our World in Data [3], which provides annual country-level statistics related to energy use and carbon emissions. Relevant features used in the study include:

- Demographic: `population`, `gdp`
- Emission sources: `coal_co2`, `cement_co2`, `oil_co2`
- Environmental indicators: `total_ghg`, `renewables_share_energy`

3.3. Data Preprocessing

The raw dataset underwent cleaning and transformation to prepare it for modeling:

- **Missing Values:** Imputed using forward/backward fill and region-wise averages.
- **Normalization:** All continuous variables were scaled using Min-Max normalization.
- **Derived Features:** New variables such as CO_2 per capita were computed.

3.4. Data Splitting Strategy

We applied a region-aware data split:

- **Training Set:** Global countries and ASEAN countries (excluding Thailand)
- **Validation Set:** ASEAN countries only (excluding Thailand)
- **Test Set:** Thailand only

This strategy ensures:

1. Real-world generalization capability
2. No data leakage (Thailand unseen during training and validation)
3. Regional pattern alignment using ASEAN countries

3.5. Model Selection and Training

We tested several models including:

- LightGBM
- Gradient Boosting
- CatBoost
- Random Forest
- XGBoost

The training was tracked using MLflow to monitor performance metrics and model parameters. Grid search and validation loss were used for hyperparameter tuning.

3.6. Evaluation Metrics

The models were evaluated using:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- R^2 Score

XGBoost showed the best performance on the test set (Thailand), confirming its robustness for this regression task.

3.7. Deployment Architecture

Figure 2 illustrates the deployment pipeline used to serve the CO_2 forecasting system. The architecture separates development and deployment responsibilities while ensuring reproducibility through Docker and Git version control.

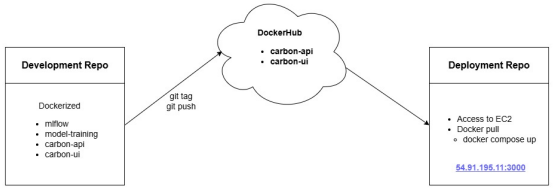


Figure 2. Deployment Pipeline using Git, DockerHub, and AWS EC2

The workflow proceeds as follows:

1. All core services—including MLflow, model training scripts, backend (carbon-api), and frontend (carbon-ui)—are developed inside a Dockerized structure within the development repository.
2. Once stable, Docker images for the API and UI are built and pushed to DockerHub using Git version tagging (`git tag`, `git push`).
3. On the deployment side, an AWS EC2 instance pulls the tagged images from DockerHub.
4. Using `docker compose up`, the application stack is launched, making the service available via `http://54.91.195.11:3000`.

This setup enables consistent deployment, easy scaling, and fast rollback if needed.

3.8. Conclusion of Methodology

By combining regionally aware training with robust deployment practices, this methodology ensures both high model accuracy and practical applicability in real-world scenarios.

4. Experiment

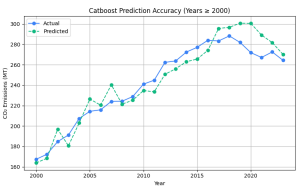
4.1. Experimental Setup

This experiment aimed to predict Thailand's CO_2 emissions using machine learning models trained on global and regional data. The goal was to determine whether models could generalize to an unseen target country by learning from other nations.

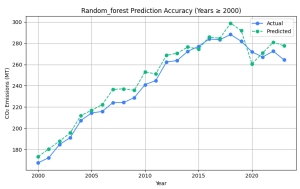
The following machine learning models were implemented and compared:

- Random Forest
- XGBoost
- LightGBM
- Gradient Boosting
- CatBoost

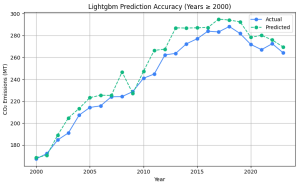
All models shared consistent preprocessing, feature engineering, and tracking using MLflow.



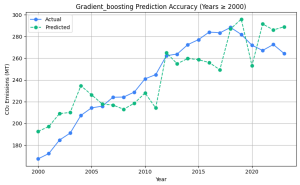
(a) CatBoost accuracy trend.



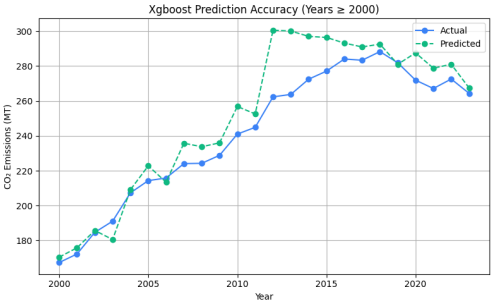
(b) Random Forest accuracy trend.



(c) LightGBM accuracy trend



(d) Gradient Boosting accuracy trend



(e) XGBoost accuracy trend

Figure 3. Comparison of training and validation accuracy trends for different models.

4.2. Data Splitting Strategy

The dataset was split based on geopolitical criteria:

- **Training Set:** Global countries and ASEAN nations (excluding Thailand)
  Purpose: Learn international and regional emission patterns.
- **Validation Set:** ASEAN countries (excluding Thailand)
  Purpose: Tune hyperparameters using regionally similar data.
- **Test Set:** Thailand only
  Purpose: Evaluate generalization to an unseen target country.

This splitting scheme mimics real-world deployment where Thailand's future data is unknown during training.

4.3. Evaluation Metrics

The following metrics were used to assess performance on Thailand's test set:

- **RMSE (Root Mean Squared Error):** Sensitive to large errors.
- **MAE (Mean Absolute Error):** Average absolute prediction error.
- **R^2 Score:** Variance explained by the model.

These metrics ensure robustness across both absolute and relative prediction errors.

5. Results

The table below summarizes the model performance on Thailand's test set:

Model	RMSE	MAE	R^2
XGBoost	**5.8318**	**3.9081**	**0.9968**
Random Forest	5.9307	4.2114	0.9967
LightGBM	6.3213	4.2731	0.9963
CatBoost	7.5339	5.0397	0.9935
Gradient Boosting	17.1723	11.8765	0.9726

Table 1. Updated model performance on Thailand's test set (1950–2023).

5.1. Key Findings

- **XGBoost** had the best overall performance, with the lowest RMSE and MAE, and the highest R^2 score.
- **Random Forest** and **LightGBM** closely followed, with comparable accuracy and excellent generalization.
- **CatBoost** achieved moderate performance, outperforming traditional Gradient Boosting.
- **Gradient Boosting** had the weakest performance, likely due to limited regularization and tuning capabilities.

5.2. Feature Importance Insights

- GDP, CO_2 per GDP, and population were among the top features across nearly all models.
- XGBoost emphasized temperature_change_from_n2o, highlighting environmental correlation with emissions.
- Different models use different importance metrics (e.g., gain for XGBoost, impurity reduction for Random Forest), explaining differences in rankings.

6. Discussion

6.1. Model Generalization

XGBoost demonstrated the highest ability to generalize Thailand's CO_2 emissions, achieving an R^2 score of 0.9968—suggesting that its boosted tree structure and regularization mechanisms were well-suited to the cross-national forecasting task. Random Forest and LightGBM also showed excellent generalization, each with R^2 values above 0.996.

6.2. Impact of Data Splitting Strategy

This experiment used a geopolitically informed split:

- Training on global + ASEAN countries (excluding Thailand),
- Validation on ASEAN countries (excluding Thailand),
- Testing on Thailand only.

This clean split simulates deployment scenarios where Thailand's data is unseen during training. The strong performance by XGBoost and Random Forest validates this approach for modeling cross-national emission trends.

6.3. Feature Importance Interpretation

The prominence of temperature_change_from_n2o in tree-based models like XGBoost and CatBoost may stem from:

- Its statistical correlation with overall emissions in training data.

- Potential biases from historical energy-environment feedback loops.

Models differ in how they calculate importance:

- **XGBoost:** Gain-based (improvement in objective function).
- **Random Forest:** Frequency of usage and impurity reduction.

Such differences can lead to varying interpretations of what features matter most.

6.4. Business Opportunities in Carbon Reduction

Beyond technical accuracy, the findings offer insight into actionable carbon reduction opportunities. The influential features identified in the models align with several growing markets in Thailand's low-carbon transition:

- **Renewable Energy Deployment and Integration:** The importance of fossil-fuel-based variables highlights business potential in solar, wind, and hydro installations to replace coal and oil dependency.
- **Energy Efficiency and Management Services:** Strong GDP and energy-related features support opportunities in smart building retrofits, industrial energy audits, and IoT-enabled energy management.
- **Carbon Markets and Offsets:** Environmental sensitivity indicators such as temperature change from N_2O emphasize the relevance of nature-based solutions, including afforestation and reforestation projects, which can generate revenue through carbon credit schemes.

These areas represent not only environmental benefits but also entrepreneurial and investment potential in Thailand's journey toward a sustainable economy.

6.5. Limitations and Future Work

- Models were not time-aware; incorporating LSTM or attention-based models may better capture long-term temporal effects.
- The analysis did not include policy shifts, global crises, or energy pricing—future versions could integrate these as exogenous signals.
- Combining satellite data, regulatory indicators, or financial market trends may enhance regional accuracy and application relevance.

7. Conclusion

This study shows that machine learning models trained on international data can effectively predict Thailand's CO_2 emissions. Among the evaluated models, XGBoost delivered the best performance (R^2 = 0.9968), demonstrating strong generalization despite the absence of Thailand in training and validation. Key predictors included GDP and environmental indicators like temperature change from N_2O.

These findings support the potential of ML-based tools not only for national emission forecasting and climate policy planning, but also for identifying business opportunities in the carbon reduction space. Areas such as renewable energy deployment, energy efficiency services, and carbon offset markets represent practical pathways where data-driven insights can guide investment and innovation.

Future work could improve forecasting accuracy by integrating time-aware models and incorporating real-world policy, climate, or economic signals.

References

[1] International Renewable Energy Agency. Renewable energy outlook: Thailand, 2021. 1

[2] R. Al-Ghezi, A. Ahmed, and P. Kumar. A machine learning framework for co emission prediction using economic indicators. *Sustainable Computing: Informatics and Systems*, 35:100753, 2022. 2

[3] Hannah Ritchie, Pablo Rosado, and Max Roser. Co and greenhouse gas emissions. *Our World in Data*, 2023. 1, 2

[4] Y. Zhao, Y. Li, and S. Wang. Carbon emission forecasting using lstm models: A case study of china. *Environmental Science and Pollution Research*, 28(10):12634–12645, 2021. 1