A generalized machine learning approach to predicting global CO2 emissions from electricity usage patterns

Opeyemi Akinniyi
University of Michigan

***Abstract*—Global energy consumption significantly contributes to greenhouse gas emissions, making it a critical factor in climate change. Accurate prediction of CO emissions from energy consumption can empower sustainable practices and inform policymaking. This study explores machine learning models, including linear regression, tree-based models, and advanced techniques like feedforward neural networks (FNN) and XGBoost, to predict global CO emissions based on energy usage patterns. Performance metrics, such as mean squared error (MSE), root mean squared error (RMSE), and R^2 score, alongside cross-validation, were employed for robust evaluation. The findings reveal that XGBoost with Optuna optimization outperforms other models, achieving the best accuracy and computational efficiency balance.**

***Index Terms*—sustainability, emissions, energy, model, optimization.**

I. Introduction

A. Background and Motivation

Global energy consumption has significantly contributed to greenhouse gas (GHG) emissions, a primary driver of climate change [1]. According to World in Data, the energy sector accounted for over 70% of global emissions in 2016. Emission estimates often rely on empirical emission factor (EF) tables, where emission (in $kgCO_2$) is calculated using the formula:

$$\text{Emission } (kgCO_2) = \text{Energy (Btu)} \times \text{EF } (kgCO_2/\text{Btu})$$

While straightforward, this approach becomes labor-intensive and time-consuming for large datasets, especially when incorporating factors like electricity generation mix, generation efficiency, and distribution losses. Moreover, the need for re-estimation as the grid mix changes adds further complexity. Developing robust machine learning models to predict CO_2 emissions can circumvent these challenges, enabling real-time insights and empowering actionable steps for sustainable practices and policymaking.

The goal of this work is to develop machine learning models, compare them, and select the best model that accurately predicts CO_2 emissions from the provided dataset.

B. Literature Review

Predicting relationships between variables using regression dates to the early 20th century, with linear regression being a cornerstone due to its simplicity and assumptions of linearity [2]. However, modern datasets often exhibit non-linear relationships and high dimensionality, challenging traditional linear methods.

Machine learning models such as random forests and gradient boosting machines have demonstrated significant improvements in handling large and complex datasets. [3] highlighted the superior accuracy of tree-based models like XGBoost for large tabular datasets. More recently, transformer-based models like BERT have proven effective for regression tasks, especially in handling high-dimensional and sequential data [4]. For energy-related CO_2 emissions:

- Gaussian Process Regression (GPR) as the best algorithm for China [5].
- support vector machines, tree ensembles, and GPR were used for a dataset covering 68 countries [6].
- The effectiveness of neural networks (FFNN, CNN) in predicting U.S. emissions was demonstrated by [7].

II. Method

A. Data Description and Preprocessing

The data was extracted from the Hugging Face database [8]. The dataset comprises annual energy consumption patterns by country, including features like electricity usage, energy sources (renewable vs. non-renewable), geographic location, land mass, and GDP per capita. The target variable is annual CO_2 emissions in metric tons.Preprocessing such as converting categorical features into numerical representations and scaling were performed.

B. Model Formulation

1) Model selection
   - **Linear Models:** Trained linear regression and lasso regression models. Linear regression is taken as the baseline.
   - **Tree-Based Models:** Trained Random forests and gradient boosting models.
   - **Advanced Models:** Feedforward Neural Network (FNN): Trained a multilayer perceptron [1] [2] [3]
2) Model Evaluation

[1]LS-1: Lasso regression with all features, LS-2: LS-1 with improved Alpha
[2]RF: 1 - Random forest, 2 - RF with grid search, 3 - RF with important features
[3]XGB: 1 - XGBoost with import features, 2- with important features and optuna search

- Mean Squared Error (MSE)
- Root Mean Square Error (RMSE)
- Coefficient of Determination (R^2)

3) **Cross-Validation:** Conducted 5-fold cross-validation for robust model evaluation. Averaged R^2 scores across folds to compare models effectively.
4) **Final Selection and Deployment:** Selected the best-performing model based on cross-validated R^2 and computational efficiency. Saved the model for deployment and real-time predictions.

III. Results

A. Linear Regression Model

Linear regression (LR) was implemented as a baseline model to predict CO_2 emissions using all the features in the dataset. The trained model had an RMSE and R^2 of 74,844 and 0.9608, respectively. while the R^2 seems like a good number, the cross-validation R^2 for the model suggested otherwise, putting the value at 0.2371. This may be a sign of overfitting; therefore, the generalization of this model is limited.

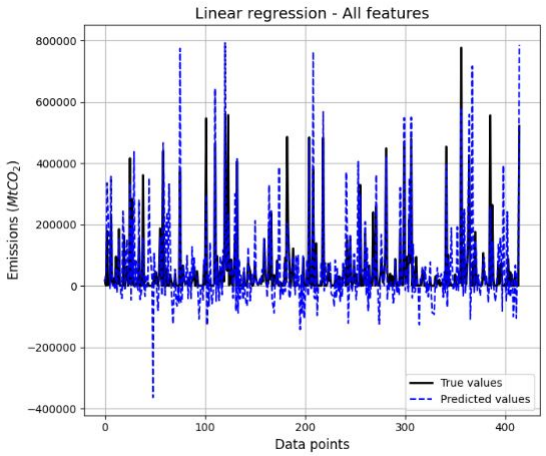

Fig. 1. Linear Model

B. Other models

The models, along with their R^2 scores, training times, and additional metrics, are presented in Fig 2.

Model	RMSE	R^2	5 fold Cross-Val R^2
LS-1	179370	0.9607	0.25173
LS-2	201259	0.9506	0.2573
RF-1	74844	0.9932	0.5597
RF-2	66499	0.9946	0.7341
RF-3	272074	0.9097	0.8538
XGB-1	47874	0.9972	0.9372
XGB-2	101466	0.9874	0.9372

Fig. 2. Table of model metrics

C. Forward Neural Networks

The initial Feedforward Neural Network (FNN-1) consists of a multi-layer perceptron with three hidden layers of decreasing dimensions: 128, 64, and 32 neurons. The Adam optimizer is used, and the learning rate (lr) and epoch were set to 0.001 and 1000, respectively. The R^2 history was tracked and plotted to check for overfitting. The trained model R^2 = 0.0918 is bad; therefore, methods like dropout and regularization were used to improve the FNN using the scenarios: FNN-2 (4 layers, R^2=0.9737), FNN-3 (4 layers and lr = 0.01, R^2=0.9972), FNN-4(3 layers, 0.2 Dropout after each layer, R^2=0.1833), FNN-5 (same as FNN-4 with lr= 0.01, R^2 = 0.9782), FNN-6 (Same as FNN-5 but 4 layers, R^2=0.9957). Compared to all models observed, the FNN improved models (i.e., drop out or changed lr) generally estimated the highest R^2 scores, making them a leading candidate for deployment. On the other hand, they took more time to process, as seen in Fig 3, compared to others. For the Models listed in Fig 2, cross-validated R2 was used in plotting Fig 3 to ensure a fair comparison with FNN scenarios.

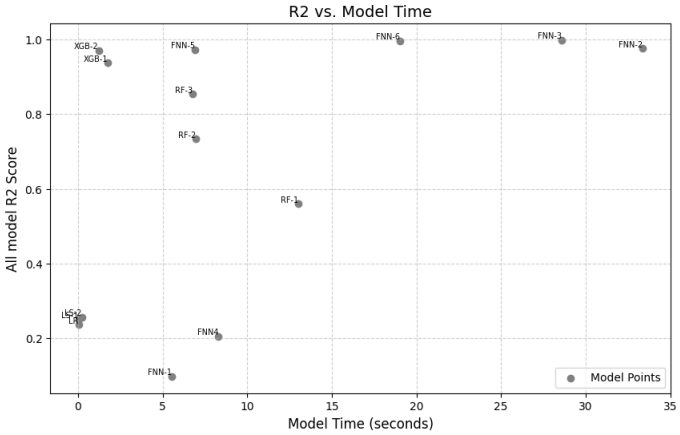

Fig. 3. Model Performance

Conclusion

The results reveal that XGBoost with Optuna optimization outperforms other models in terms of predictive accuracy and computational efficiency and, therefore, is chosen and saved for future deployment.

Selected References

1 International Energy Agency (IEA), "World energy balances," 2021. [Online]. Available: https://www.iea.org

3 T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

4 Z. Liu, et al., "Predicting energy consumption using machine learning: A review," *Energies*, vol. 11, no. 9, p. 2434, 2018.