Predicting National CO_2 Emissions: A Comparative Study of Linear Regression, XGBoost, and Time-Series Models

Serena Chen, Ryan Cho, Sereena Gill, Jack Iorio, Unce Shahid

1. Abstract

As global CO_2 emissions reach record highs, predictive modeling has become a critical tool for informing climate policy and sustainability planning. In this project, we use country-level data from the Our World in Data (OWID) dataset to forecast future CO_2 emissions based on historical emissions trends and key economic and energy indicators. We implement and compare multiple machine learning models—including Linear Regression, XGBoost, and ARIMAX—evaluating performance using RMSE and MAPE metrics. To improve transparency and interpretability, we apply SHAP to identify the most influential features driving emissions predictions. Our results show that autoregressive models such as ARIMAX outperform other approaches by effectively capturing temporal dependencies in national emission trends. These findings highlight the value of integrating time series analysis with interpretable ML to support evidence-based environmental decision-making.

2. Background

Global carbon dioxide (CO_2) emissions are at unprecedented levels, driving critical climate challenges such as rising temperatures, extreme weather events, and widespread environmental disruption. In response, this project aims to develop predictive models that estimate future CO_2 emissions on a country-by-country basis. Such forecasting is vital for identifying high-risk nations and informing proactive climate policies, especially in rapidly industrializing regions where emissions are accelerating. While prior studies have utilized machine learning models (such as Random Forest, Support Vector Regression, and Neural Networks) to predict emissions within single countries, there remains a notable gap in scalable, interpretable models that generalize across multiple nations. By addressing this limitation, our project seeks to contribute actionable insights to global sustainability efforts.

3. Methods

3.1 Data Collection and Preprocessing

Our analysis utilized a country-level dataset from Our World in Data (OWID), which provides annual records on CO_2 emissions, energy use, and macroeconomic indicators. We restricted the dataset to individual countries, excluding aggregate regions such as "World" and

"Asia" to ensure that each row represented a distinct national observation. The data was further filtered to cover the period from 1990 to 2023, providing a consistent and modern time frame.

We selected a targeted set of features relevant to emissions prediction, including GDP, population, primary energy consumption, and sector-specific CO_2 emissions (coal, oil, gas, and cement). This selection was guided by domain knowledge and the availability of complete records. To maintain data quality, rows with missing values in any of the selected features were removed.

All numeric features were standardized using StandardScaler, transforming them to have zero mean and unit variance. This scaling was critical for ensuring that models like Linear Regression and XGBoost, which are sensitive to feature magnitudes, performed optimally. The data was then divided into training (1990–2014) and testing (2015–2023) subsets, using a chronological split to ensure that the model was evaluated on unseen future data.

3.2 Linear Regression and XGBoost

To establish predictive baselines for national CO_2 emissions, we implemented both Linear Regression and XGBoost using the country-level dataset from Our World in Data. These models were selected to provide contrast between a simple, interpretable method and a more complex, nonlinear one. Linear Regression served as our baseline due to its ease of implementation, direct coefficient interpretability, and strong performance on structured data. XGBoost, on the other hand, is a gradient-boosted tree model designed to capture nonlinearity and higher-order interactions between features.

We trained both models on scaled input features including GDP, population, primary energy consumption, and sector-specific emissions (coal, oil, gas, etc.), with the target variable being total national CO_2 emissions. To evaluate performance, we used two standard regression metrics: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The data was split by year, with the training set comprising data from 1990 to 2014, and the test set covering 2015 to 2023.

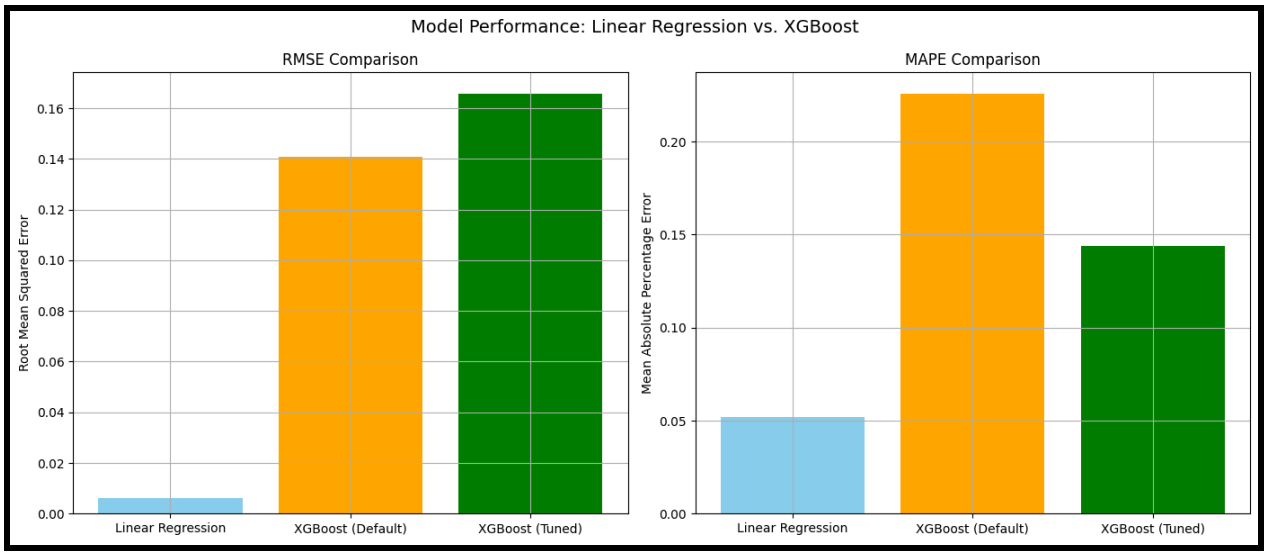

Figure 1: RMSE and MAPE scores for Linear Regression, default XGBoost, and tuned XGBoost models.

Results indicated that Linear Regression outperformed XGBoost across both evaluation metrics, as shown in **Figure 1**. Linear Regression achieved an RMSE of 0.006 and a MAPE of 5.2%, whereas default XGBoost exhibited a significantly higher MAPE of 22.6%. Even after hyperparameter tuning via GridSearchCV—testing combinations of learning rate, tree depth, and number of estimators—XGBoost performance improved only modestly, yielding a best MAPE of 14.4%. This suggests that the underlying relationship between macroeconomic and emissions indicators is largely linear, and that the additional complexity introduced by XGBoost may have led to overfitting or diminishing returns.

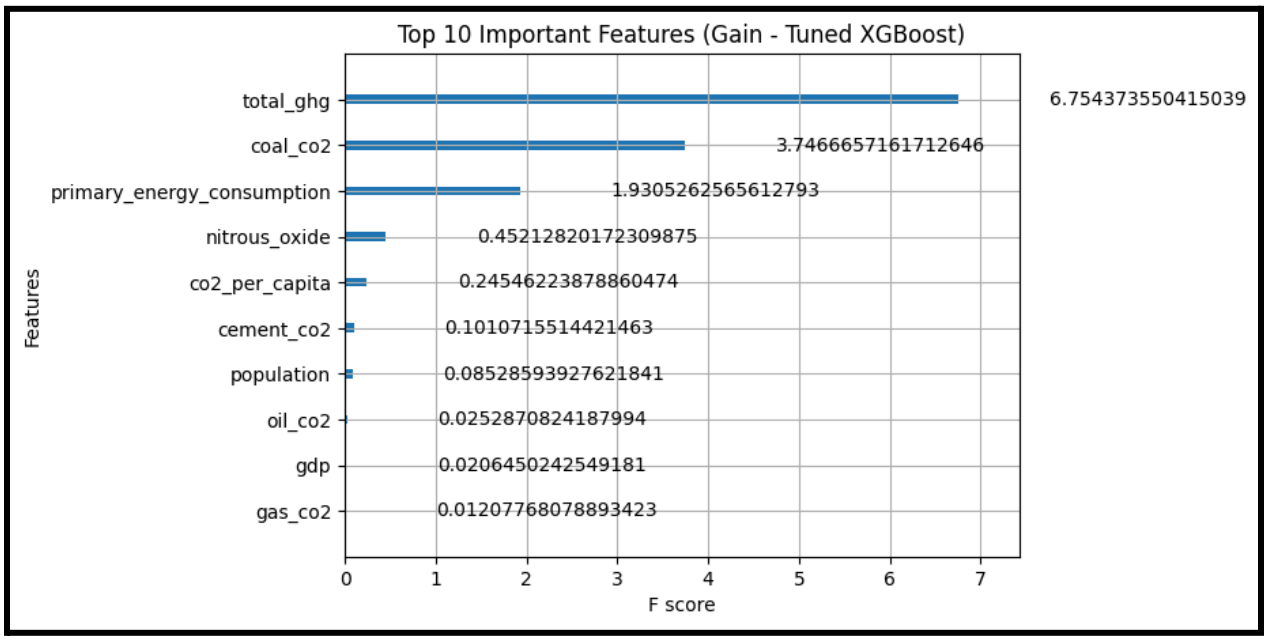

Figure 2: Gain-based feature importance from the tuned XGBoost model.

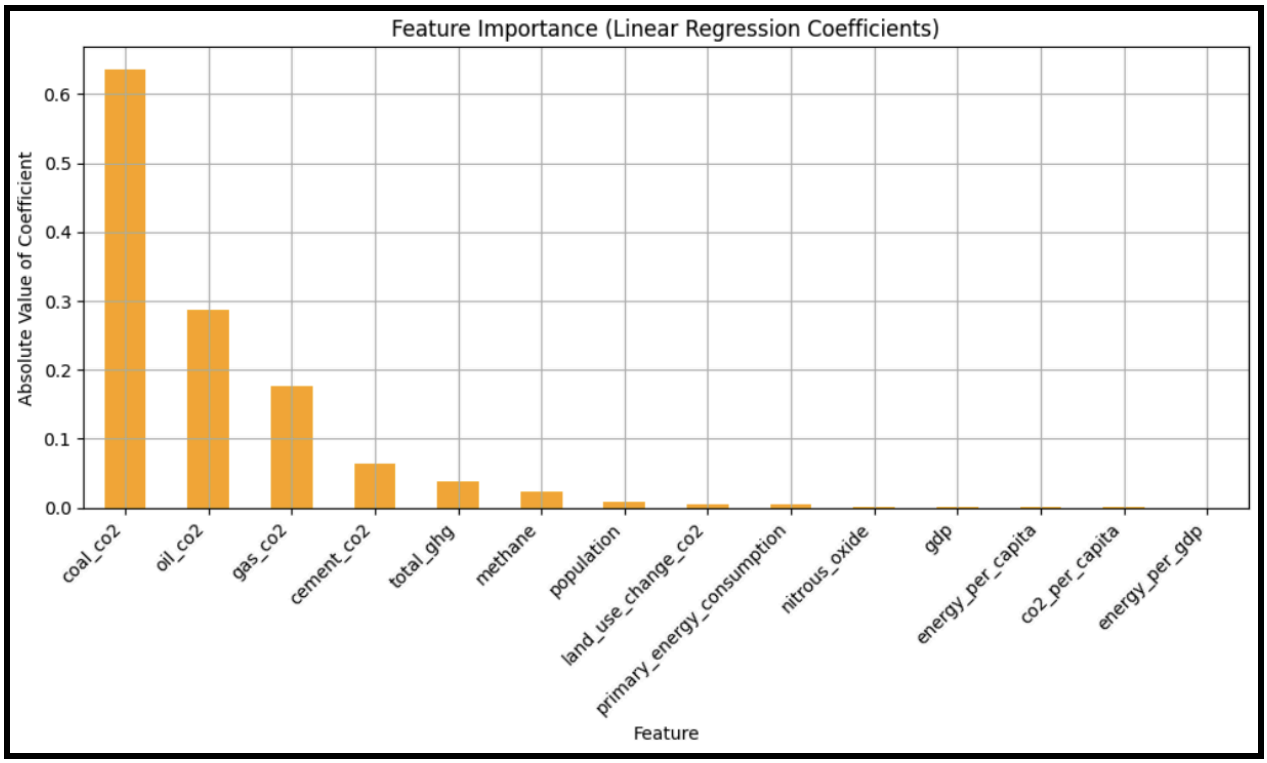

Figure 3: Gain-based feature importance from the tuned XGBoost model.

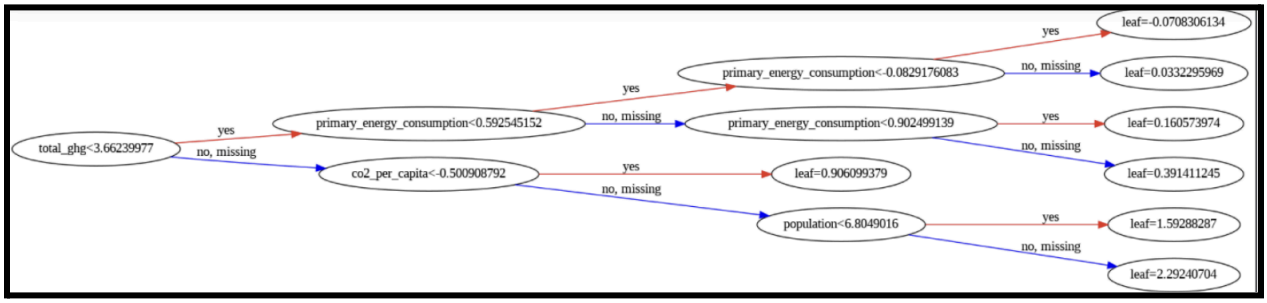

Figure 4: Tuned XGBoost's Decision Pathways

To further interpret model behavior, we examined feature contributions using both SHAP and gain-based feature importance from the tuned XGBoost model. As shown in **Figure 2**, the most influential predictors of national emissions were energy-related variables, including total greenhouse gases (total_ghg), coal emissions (coal_co2), and primary energy consumption. These findings are consistent with domain knowledge and reinforce the view that fossil fuel combustion remains the dominant driver of CO_2 output at the national level.

Additionally, we examined the feature importance from the Linear Regression model by analyzing the absolute values of the model's coefficients. As shown in **Figure 3**, coal emissions (coal_CO2) had the largest absolute coefficient, followed by oil_co2 and gas_co2. These fossil fuel indicators dominated the model, reinforcing the climate literature that emphasizes fossil fuels as a primary source of emissions. In addition, the simplicity and transparency of a coefficient-based interpretation also demonstrates one of the key advantages of a linear model.

To enhance transparency of our methods, we visualized a *single* decision tree from the tuned XGBoost model as shown in **Figure 4**. This tree illustrates how the model sequentially splits the data on key features such as: total_ghg, primary_energy_consumption, and co2_per_capita to make emissions predictions. The way this works is at the root node, the model begins by evaluating whether a country's total greenhouse gas emissions (total_ghg) falls below a certain threshold (≈ 3.66). This early split shows the variable's importance in how it shapes model predictions. Subsequent branches examine other thresholds (such as energy consumption and per capita CO2). The tree-based structure allows the model to capture more nuanced and nonlinear interactions between features. Furthermore, this visualization provides an intuitive way to trace the logic behind each prediction. It is important to note this is just one tree from the full XGBoost ensemble, which typically consists of hundreds of trees that differ in root nodes, branch structure, and leaf values. These differences enable the model to capture diverse patterns across the dataset, and their collective output produces the final emissions prediction.

Overall, the performance of Linear Regression illustrates that well-engineered features can yield accurate results without requiring high model complexity. XGBoost, while powerful in theory, underperformed in this context—likely due to the linear nature of the relationships and the relatively low sample size compared to the feature space.

3.3 ARIMAX

Given that the OWIF data is represented in time-series format, we decided to use a time-series model. As a typical ARIMA model is unable to receive features as inputs (only uses the target feature), we chose to run an ARIMAX model instead. The ARIMAX model uses the same set of features that Linear Regression and XGboost use, allowing for easy comparisons. The ARIMAX model takes in the hyperparameters 3, 1, and 3 to consider multiple years when making predictions and only differencing once. The data is split into training and test data based on years. An 80/20 split is formed, where data from 1990 to 2015 is being trained on, while data from 2016 to 2022 is being tested on. There is no validation set as there are too few years as is for training and testing, so hyperparameter tuning would not yield useful results.

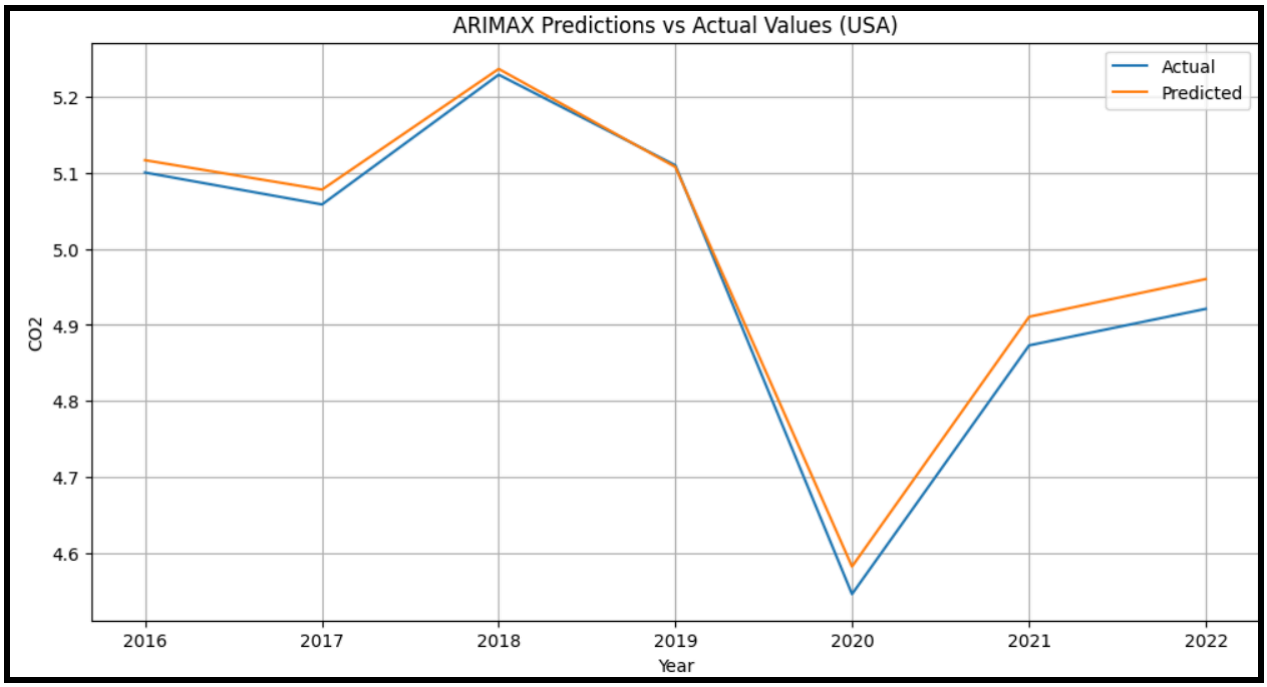

Figure 5: ARIMAX predictions on USA

Figure 5 is a line plot comparison between the predictions of the ARIMAX model and the test data on USA values in specific. As can be seen from the diagram, the model is able to learn the general shape of the United State's carbon dioxide emissions trend very well. The ARIMAX model achieves an RMSE score of .027 and a MAPE score of .005 in this test.

One limitation of time-series models is that they can only be trained on data from a singular country at a time. Time-series models must be fed continuous data, where multiple country data is no longer continuous. One way to overcome this issue is to use neural networks and to feed in different instances of ARIMAX models to attempt to achieve a generalized model to predict carbon dioxide emissions for various countries. However, such a model would require an immense amount of data, of which the data we have may already not be sufficient enough for simple model implementation purposes. Instead, we decided to implement a new ARIMAX model for each individual country and aggregate their performance metrics. This multitude of ARIMAX models runs on the same philosophy as before with an 80/20 train/test split and a hyperparameter set of 3,1,3. Some of the results can be seen in **Table 1**.

	Country	RMSE	MAPE	R2
0	Afghanistan	0.000005	0.000013	0.999969
1	Albania	0.000029	0.000073	0.984836
2	Algeria	0.000271	0.001490	0.999065
3	Argentina	0.001554	0.009552	0.982025
4	Armenia	0.000002	0.000008	0.999990
...	...	...	...	...
96	Uruguay	0.000002	0.000004	0.999994
97	Uzbekistan	0.000988	0.004761	0.982281
98	Venezuela	0.005118	0.019083	0.965502
99	Vietnam	0.002155	0.042164	0.998310
100	Yemen	0.000363	0.001100	0.393924

101 rows × 4 columns

Table 1: A table of RMSE, MAPE, and R2 values for each country.

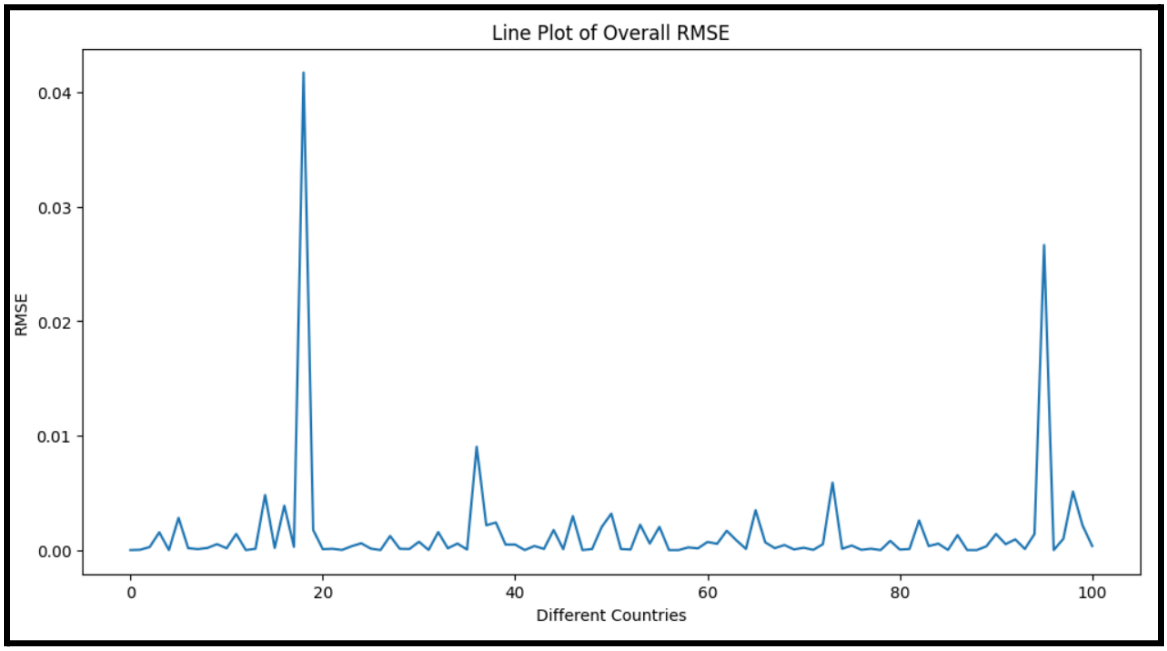

Figure 6: Plot of the varying RMSE scores per country.

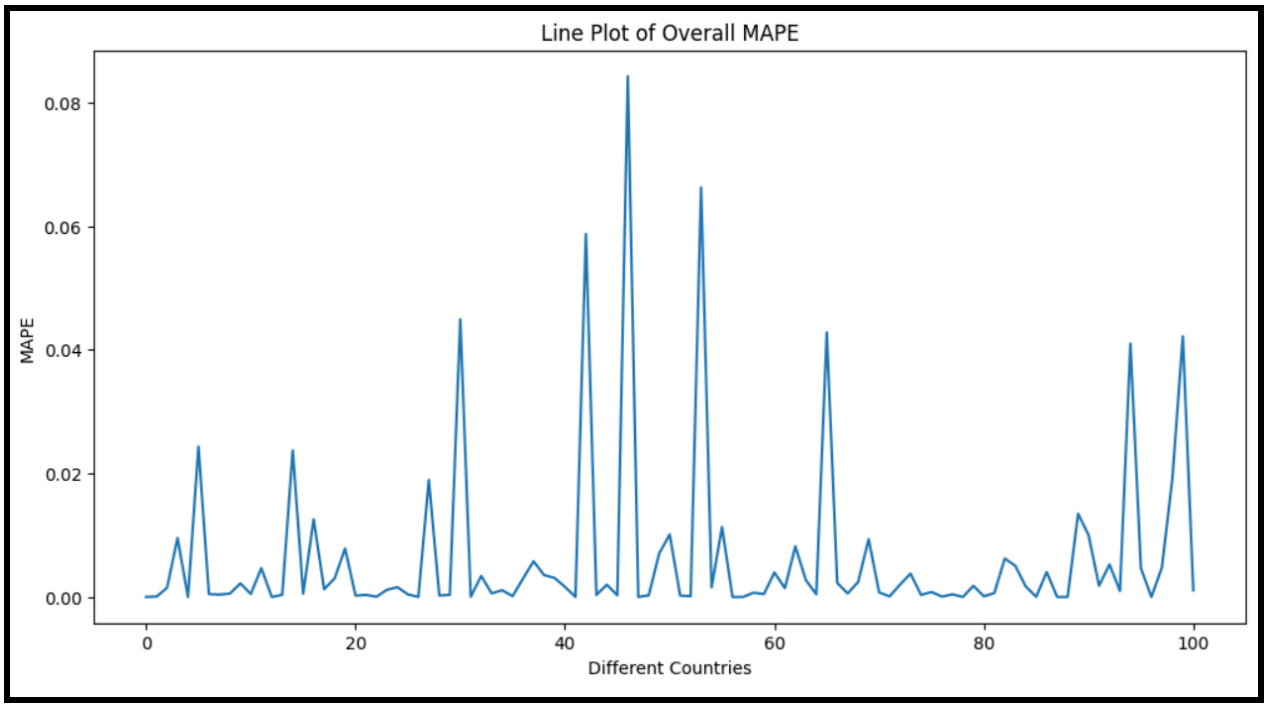

Figure 7: Plot of the varying MAPE scores per country.

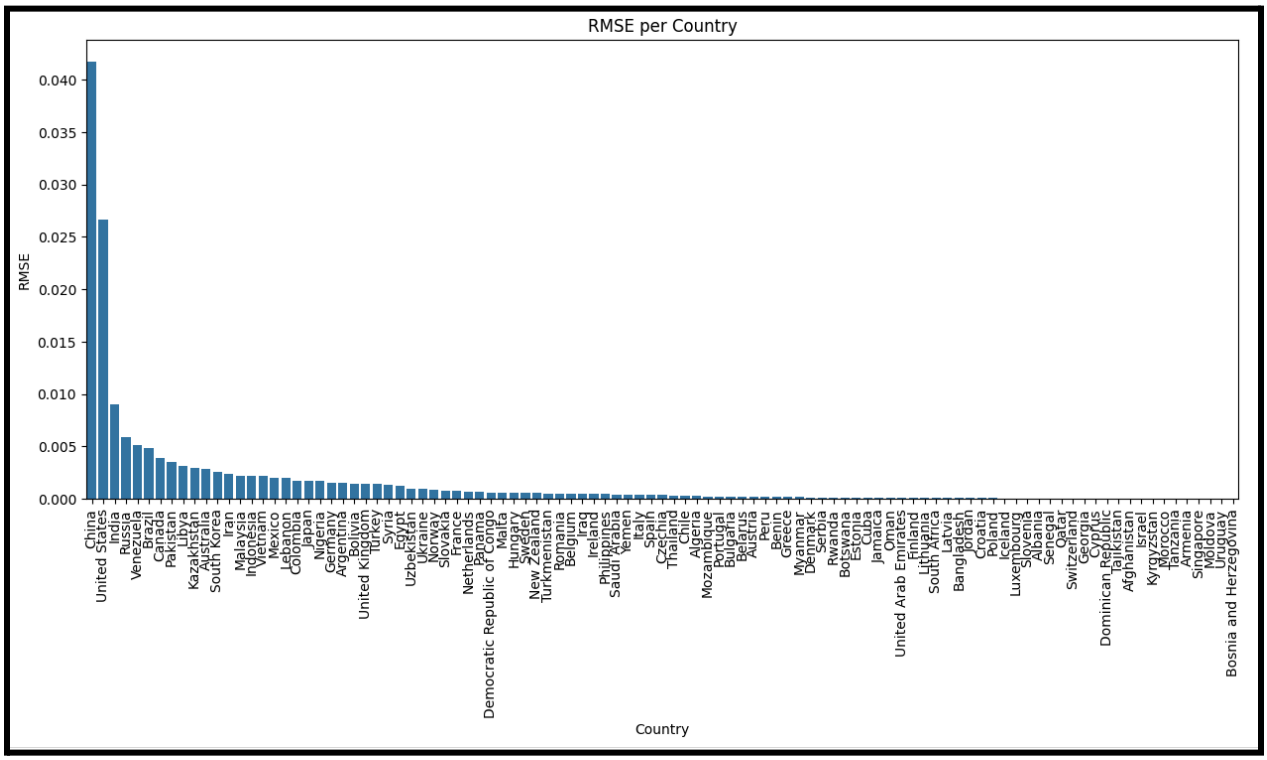

Figure 8: Bar plot of RMSE scores for each country sorted in descending order.

Figure 6 plots the various RMSE scores per country, **Figure 7** plots the various MAPE scores per country, and **Figure 8** plots the RMSE scores with country names in descending order. As can be seen by **Figure 6** and **Figure 7**, RMSE and MAPE scores are generally low for

most countries, but end up spiking for a select few countries. **Figure 8** allows us to identify these countries. The biggest RMSE scores exist in the following two countries: China, and the United States. However, the rest of the RMSE scores are very low. 95% of the countries (96 out of 101 total countries) have an RMSE score that is less than 0.005. Furthermore, the average RMSE value across all countries' metrics is 0.0016, with the average MAPE being 0.0067. For a model that attempts to generalize carbon dioxide emissions, these scores are rather good. As such, we believe the ARIMAX models to be a generally good predictor of carbon dioxide emissions that can be applied to individual countries.

3.4 Evaluation Metrics

To assess the performance of our models, we used two common regression evaluation metrics: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE).

The RMSE measures the square root of the average squared differences between predicted and actual values. It is expressed in the same units as the target variable, CO_2 emissions, which makes it easy to interpret. Additionally, what makes this metric useful is the fact it penalizes larger errors more heavily due to the squaring operation, which is particularly useful when significant deviations from the actual value are more important to consider. A lower RMSE indicates a model that more accurately captures emission trends.

Conversely, the MAPE measures the average absolute difference between predicted and actual values as a percentage of the actual values. Unlike RMSE, it is unitless and has an easier interpretation as it can be compared more easily to different countries. A lower MAPE suggests better model accuracy in relative terms, for example a MAPE of 12% suggests predictions were off by an average of 12% from the actual emission values.

Both RMSE and MAPE provide both absolute and relative perspectives on model performance, helping us identify models that are not only accurate overall but also are proportionally consistent across the different levels of emissions. **Table 2** gives more detail into the metrics used.

Root Mean Squared Error	$\sqrt{\frac{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}{n}}$	y_i = The actual value of the target variable for observation *i* $\widehat{y}_i$ = The predicted value of the target variable for observation *i* n = Total number of observations

Mean Absolute Percentage Error	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert A_i - F_i\rvert}{A_i}$	A_i = The actual value F_i = The forecast value n = Total number of observations

Table 2: Evaluation Metrics

3.5 Interpretation and Explainability

A notable problem in past models and literature related to the prediction of CO2 emissions by country is the lack of explainability and interpretability in the processes and results of the models. This is primarily due to the use of black box methods and a lack of feature explanations, making it difficult to understand what features are most influential in model predictions of CO2 emissions. We attempt to address this by building various models that allow us to extract feature importances, SHAP in the case of the linear regression and XGBoost models, and feature coefficients in the case of the ARIMA and ARIMAX models.
For the linear regression and XGBoost models, SHAP was used to explore feature importances for each model.

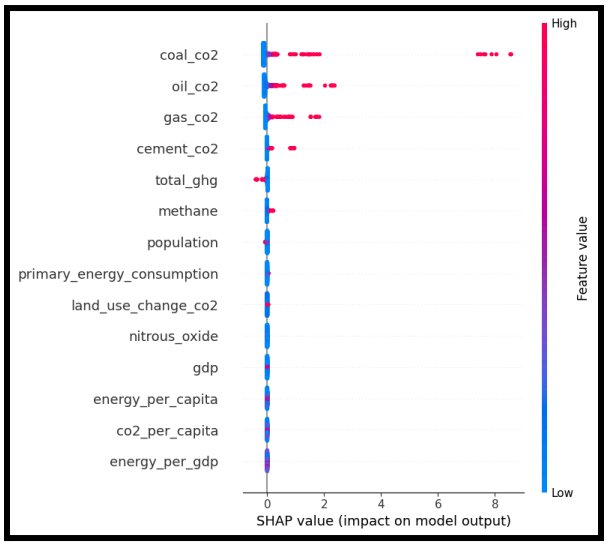

Figure 9: SHAP Summary Plot for Linear Regression model

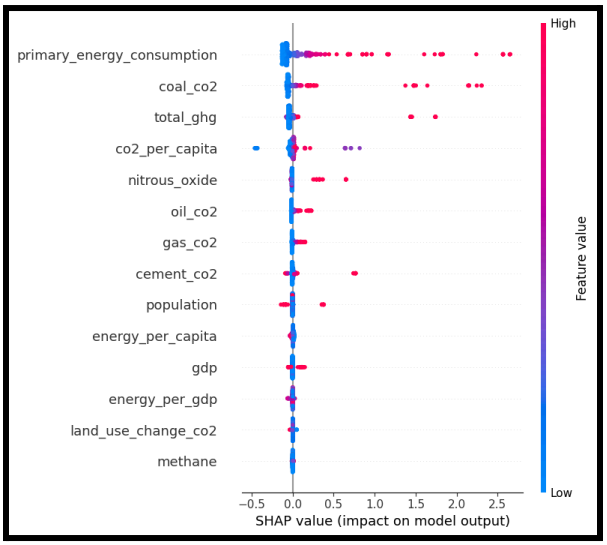

Figure 10: SHAP Summary Plot for XGBoost model

Figure 9 displays the overall SHAP summary plot for the linear regression model. The summary plot indicates that the use of fossil fuels (as shown by coal_co2, oil_co2, gas_co2) has a highly positive impact on the model's predictions for a country's given CO2 emissions. **Figure 10** displays a similar behavior, but appears to account for the country's energy consumption in addition to its fossil fuel usage. These summary plots allow us to better understand the feature

importances of each individual model, giving us the information necessary to adjust and interpret these models for future use in both modelling and analysis.

For the ARIMA and ARIMAX models, feature coefficients were produced to aid with the explainability of the model.

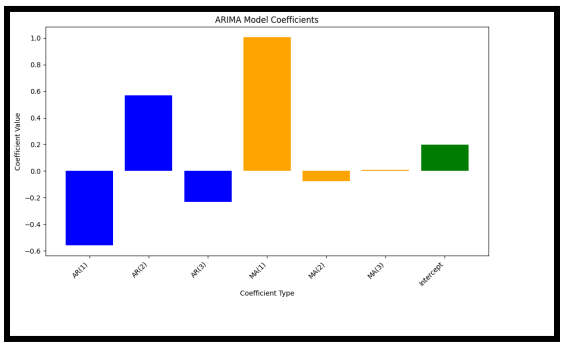

Figure 11: Feature coefficients for ARIMA model

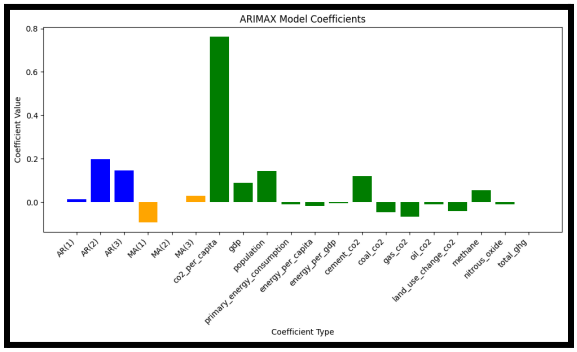

Figure 12: Feature coefficients for ARIMAX model

Figure 11 displays the feature coefficients produced for the ARIMA model. The AutoRegressive terms appear to oscillate, indicating that if CO2 emissions were high in the previous period, the model predicts they will drop in the next period and vice versa. The Moving Average terms show a strong effect from the previous prediction error, meaning that any recent fluctuations in error have a heavy influence on the model's predictions. The model's intercept indicates the baseline level of CO2 emissions for the model. **Figure 12** displays the feature coefficients produced for the ARIMAX model. The AutoRegressive terms indicate that the past CO2 emissions has a positive impact on predictions for current emissions. The Moving Average terms indicate that recent forecast errors have a slightly negative influence on model predictions. In the exogenous variables, co2_per_gdp appears to have a strong positive impact on the model's predictions. Given the nature of this variable as a likely byproduct of processes that produce CO2, it is likely that this variable attempts to correct other variables like oil_co2 to avoid double-counting the influence of the variables. Otherwise, features like primary_energy_consumption and population seem to have a positive impact on the model's predictions. These coefficients give us greater insight into how past predictions, prediction errors, and the features of these models influence the models' predictions in the present, which is crucial for being able to enhance the model and explain its results.

4. Results Summary

In **Table 3**, we summarized the performance of all models tested, comparing their RMSE, MAPE, and key features. The AutoRegressive model achieved the strongest overall performance, with the lowest RMSE (0.000157) and MAPE (0.67%), indicating high accuracy in capturing temporal trends for national CO_2 emissions.

To provide a comparative context, we included two benchmark models from the literature depicted in the green-shaded rows. These models predict CO_2 *per capita* using features identified through SHAP (Shapley Additive Explanations), as the original study did not report model coefficients. While the XGBoost model from Le et al. achieved competitive MAPE, it under-performs the ARIMAX model in RMSE, which reinforces the advantages associated with temporal structure in forecasting models.

Model	RMSE	MAPE (%)	Key Features Used	Notes
Linear Regression	0.006	5.20	1) Coal CO_2 2) Oil CO_2 3) Gas CO_2	Predicts total CO_2 emissions (not per capita); features are fuel-specific
XGBoost	0.141	22.60	1) Total Greenhouse Gasses 2) Primary Energy Consumption 3) Coal CO_2	Predicts total emissions; feature set includes total GHG + energy indicators
Tuned XGBoost	0.166	14.40	1) Total GHG 2) Coal CO_2 3) Primary Energy Consumption	Hyperparameters optimized; total emissions target; key features reordered by importance
AutoRegressive Model	0.00157	0.67	1) GDP 2) Population 3) Primary Energy Consumption	Time series model using lagged emissions; predicts total CO_2 emissions
Linear Regression *	0.14	9.26	1) Population 2) GDP	Predicts CO_2 per capita (tons/person); values from Le et al. (2024)
XGBoost *	0.019	0.51	1) Population 2) GDP	*Predicts CO_2 per capita (tons/person); values from Le et al. (2024)*

Table 3: Model Performance Comparison Across CO_2 Emission Forecasting Methods

* Rows highlighted in green represent benchmark results from Le et al. (2024), included for comparison. Feature importances for these models were identified using SHAP (Shapley Additive Explanations), as the original paper did not report model coefficients.

5. Discussion

Our results indicate that the AutoRegressive models work best with the data, likely due to its time-series forecasting capabilities. When considering the coefficients for the ARIMA and ARIMAX models, the models appear to weigh many more of the variables potentially related to CO2 emissions. However, there does appear to be some limitations in how the model is able to interpret the variables while accounting for multicollinearity. But as one of the main goals of this implementation was to predict future CO2 emissions, the low RMSE and MAPE scores for the AutoRegressive models indicate they definitively outperform the other models even when accounting for these limitations.

If we were to continue this project, there are several directions that could meaningfully enhance policy relevance and the robustness of our emission predictions. First, we could expand the data that is used to more regional data or perhaps data by specific city. In addition, we could consider using an LSTM-based recurrent neural network to more effectively model long-term temporal dependencies with emission trends. Lastly, we could consider more causal inference frameworks to move beyond correlation-based prediction and to gauge a better understanding of what the drivers of emissions are. Incorporating causal approaches would not only improve the model's interpretability but also give more actionable insights for policymakers to seek more design effective targeted climate strategies.

Our study aligns with existing literature that applies machine learning techniques for CO_2 emissions forecasting. Specifically, our use of Linear Regression, XGBoost, and ARIMAX models reflects common approaches found in the literature. Jiang and Ma (2020) employed a grey multivariable model to forecast China's CO_2 emissions, integrating economic and energy indicators similar to our use of GDP and energy consumption. While their model is tailored for China, our ARIMAX model demonstrates the scalability of time-series models across multiple countries.

Khoshnevis Yazdi and Shakouri (2019) utilized an optimized artificial neural network (ANN) model, enhanced by correlation and principal component analysis, to predict greenhouse gas emissions. Our findings suggest that simpler models like Linear Regression can achieve comparable accuracy when key features are properly selected, highlighting the importance of interpretability.

Nguyen and Le (2022) combined machine learning techniques, including XGBoost, with time-series models to forecast energy consumption and CO_2 emissions in Vietnam. Their focus on time-series models aligns with our observation that ARIMAX outperforms other models in capturing temporal dependencies. Additionally, their use of SHAP for feature importance mirrors our approach, reinforcing the value of explainable AI.

Collectively, these studies emphasize the value of integrating economic and energy indicators for emissions forecasting, which our study builds on. Our research extends this by

demonstrating the scalability of ARIMAX across countries and balancing model complexity with interpretability.

6. References

a. Jiang, P., & Ma, X. (2020). *Forecasting Chinese CO_2 emissions from fuel combustion using a novel grey multivariable model*. Journal of Cleaner Production, 244, 118640. https://doi.org/10.1016/j.jclepro.2019.118640

b. Khoshnevis Yazdi, S., & Shakouri, B. (2019). *Forecasting greenhouse gas emissions using an optimized artificial neural network model based on correlation and principal component analysis*. Journal of Cleaner Production, 218, 886–896. https://doi.org/10.1016/j.jclepro.2019.02.043

c. Nguyen, T. H., & Le, T. T. (2022). *Forecasting energy consumption and carbon dioxide emission of Vietnam by prognostic models based on explainable machine learning and time series*. Energy Reports, 8, 2787–2801. https://doi.org/10.1016/j.egyr.2022.01.107

d. Ritchie, H., Roser, M., & Rosado, P. (2024). *CO_2 and Greenhouse Gas Emissions*. Our World in Data. https://ourworldindata.org/co2-and-greenhouse-gas-emissions