Elements of AI ML

Assignment 2 - (Research Paper)

Title :

Predicting CO2 Emissions Per Capita Over Time: A Machine Learning Approach for Trend Analysis

Name : Ishita Pradhan

Batch : 11

Roll no : R2142230350

Sap Id : 500119414

Title:

Predicting CO2 Emissions Per Capita Over Time: A Machine Learning Approach for Trend Analysis

Abstract

Understanding country contributions to climate change and evaluating the effectiveness of environmental programs depend heavily on **CO2 emissions per capita**. Using only historical data from 1990 to 2018, this study **forecasts CO2 emissions per capita for future years** using machine learning techniques. **Mean Absolute Error (MAE), Mean Squared Error (MSE)** and **R2 score** are three common regression measures that we use to assess the effectiveness of three machine learning models—**Linear Regression, Random Forest Regressor** and **Gradient Boosting Regressor**—in predicting emissions. Despite the challenges in predicting CO_2 emissions, the results show that the **Linear Regression** model performs the best in terms of predictive accuracy, with the lowest MAE and MSE and the highest R^2 score compared to the **Random Forest** and **Gradient Boosting** models. Additionally, **Explainable AI (XAI)** techniques, specifically **SHAP** and **LIME**, are used to provide transparency and interpretability of the predictions. These methods help ensure that the machine learning models' decision-making processes are transparent and understandable, providing valuable insights for policymakers and emphasize the utility of machine learning techniques in forecasting future emissions, even with limited data. By providing actionable insights into emission trends, this study helps inform data-driven climate policies aimed at **mitigating climate change** and achieving sustainability targets.

1. Introduction

1.1 The Global Challenge of CO2 Emissions and Climate Change

The increase in **CO2 emissions** is widely recognized as one of the primary drivers of **global warming** and climate change. CO2, along with other greenhouse gases, traps heat in the Earth's atmosphere, leading to rising global temperatures, changing weather patterns, and a variety of environmental disruptions. As a result, understanding the sources and trends of CO2 emissions is essential for combating climate change and reducing the human impact on the environment.

CO2 emissions per capita is an essential metric because it allows for a more equitable assessment of a country's contribution to climate change, adjusting for population size. While a few countries may contribute the most in total emissions, their emissions per capita might be much lower than those of highly industrialized nations. Hence, emissions per capita provide a clearer picture of a nation's individual carbon footprint.

Forecasting CO2 emissions per capita is a complex task, as emissions are influenced by numerous variables such as population growth, economic expansion, energy consumption, industrialization, and technological development. More recently, machine learning (ML) methods have been increasingly utilized to tackle forecasting challenges across various domains, including climate science. These methods can help identify hidden patterns and provide more accurate predictions compared to traditional statistical methods.

This study aims to predict CO2 emissions per capita over time using machine learning techniques based on historical emissions data. The dataset available for this study covers emissions data from **1990 to 2018**, providing a significant time span to analyze trends and patterns. Unlike other studies that incorporate a range of socio-economic and political variables, this study focuses solely on **historical emissions data**, which provides a simplified but challenging problem for forecasting.

To address this challenge, this research applies three machine learning models: **Linear Regression**, **Random Forest Regressor**, and **Gradient Boosting Regressor**. Each model has different strengths, and this study aims to compare their performance in predicting CO2 emissions per capita.Additionally, we incorporate **Explainable AI (XAI)** to ensure that the results are transparent, interpretable, and actionable for decision-makers.

1.2 Research Objectives

- To forecast CO2 emissions per capita using machine learning regression models based on historical data.
- To compare the performance of Linear Regression, Random Forest, and Gradient Boosting in terms of predictive accuracy.
- To employ **Explainable AI (XAI)** techniques (SHAP and LIME) to enhance the interpretability of the machine learning models and provide insights into the factors driving the predictions.

1.3 Research Significance

This study is significant because it demonstrates how machine learning techniques may be applied to predict future CO2 emissions, one of the most pressing issues facing the globe. By focusing on past emissions data, the study shows how machine learning algorithms can spot important trends and forecast future emissions with remarkable accuracy. There are important implications for policymakers: accurate forecasting helps governments to understand the long-term impacts of their decisions, set realistic goals for decreasing emissions, and implement proactive climate policies.

2. Literature Review

2.1 Machine Learning in Environmental Modeling

The use of machine learning in environmental science has gained significant traction in recent years, particularly in the context of forecasting environmental variables like CO2 emissions, temperature rise, and energy consumption. While traditional statistical models such as **ARIMA (Autoregressive Integrated Moving Average)** have been used for time series forecasting, these methods often assume linear relationships between variables. In contrast, machine learning models can handle **non-linear relationships**, making them more suitable for complex environmental data.

Machine learning algorithms such as **Random Forests** and **Gradient Boosting Machines** have become increasingly popular due to their ability to process large datasets and uncover hidden patterns. These models can capture intricate relationships between variables, allowing for more accurate predictions in cases where traditional methods struggle. **García et al. (2019)** demonstrated that **Gradient Boosting** outperformed simpler linear models in predicting CO2 emissions and energy consumption, thanks to its capacity to model complex, non-linear relationships.

Another important advantage of machine learning models is their ability to work with high-dimensional datasets. For example, **Bocchi et al. (2020)** applied **Random Forests** to predict climate-related time series, showcasing its effectiveness in handling multivariate datasets. However, as this study focuses on **single-variable forecasting**, the key challenge is to predict emissions based solely on historical CO2 emissions data, without the inclusion of additional socio-economic variables.

2.2 Forecasting CO2 Emissions: A Historical Perspective

Many studies have analyzed the long-term trends in global CO2 emissions, identifying the major drivers behind emissions growth. Anderson et al. (2019) noted that the global emissions trajectory has been heavily influenced by industrialization, technological advancements, and energy consumption. CO2 emissions have risen steadily over the last century, primarily due to increased use of fossil fuels, particularly coal, oil, and natural gas. However, economic downturns, such as the **2008 financial crisis**, have temporarily reduced emissions due to decreased industrial output and energy consumption. This cyclical nature of emissions underscores the complexity of forecasting emissions, as economic growth and recession cycles play a significant role.

In addition to long-term growth, research by **Pfeiffer et al. (2018)** also highlighted that emissions fluctuations are often linked to major global events, such as energy crises or technological advancements. For example, **technological shifts towards cleaner energy** sources, such as solar and wind, have the potential to significantly reduce emissions in the future. However, these shifts take time to implement, and they may not immediately offset the growing demand for energy in many industrialized and developing nations.

2.3 The Importance of Time Series Forecasting for CO2 Emissions

Time series forecasting is particularly relevant in the context of CO2 emissions because the underlying trends are often **seasonal** or **cyclical**. While the long-term trend in global emissions is generally upward, short-term fluctuations—driven by factors such as **economic activity**, **technological advancements**, and **policy changes**—also influence emissions patterns. Accurately predicting future emissions involves not only identifying the long-term growth trend but also accounting for these short-term variations.

Traditional time series methods like **ARIMA** or **Exponential Smoothing** can be effective when the data exhibits linear trends or seasonal patterns. However, these methods struggle when data is subject to more complex, non-linear relationships. This is where machine learning models such as **Random Forest** and **Gradient Boosting** excel, as they can automatically learn from data and adapt to changing patterns over time. The potential of these models to handle non-linearities makes them particularly promising for forecasting CO2 emissions, where trends can be influenced by numerous unpredictable factors.

2.4 Previous Applications of Machine Learning in CO2 Emissions Forecasting

Machine learning has already demonstrated its potential in predicting CO2 emissions. For example, **García et al. (2019)**used Gradient Boosting to predict CO2 emissions and energy consumption, showing that boosting methods provide superior results compared to linear regression models. Similarly, **Bocchi et al. (2020)** applied Random Forests to climate-related time series, finding that ensemble models outperformed traditional methods.

These studies provide a strong foundation for this research, as they demonstrate the efficacy of machine learning models in environmental forecasting. However, this study goes a step further by focusing specifically on **CO2 emissions per capita**, using only historical emissions data and comparing the performance of three different machine learning models.

2.5 The Role of Explainable AI in Environmental Forecasting

As machine learning models grow more complex, the need for interpretability and transparency has become crucial, especially in fields like environmental forecasting, where model predictions directly influence policy and decision-making. **Explainable AI (XAI)** addresses this by providing methods to explain how models arrive at their predictions, fostering trust among stakeholders.

Two widely-used XAI techniques in environmental modeling are **SHAP (SHapley Additive exPlanations)** and **LIME (Local Interpretable Model-agnostic Explanations)**. SHAP, based on game theory, assigns a value to each feature representing its contribution to a prediction, offering both global and local insights into the model's behavior. LIME, a model-agnostic method, explains individual predictions by approximating the complex model with simpler interpretable models locally around each data point.

These techniques have been applied in environmental fields like emissions forecasting and energy consumption prediction. For example, SHAP helps identify which factors, such as policy shifts or energy use, most affect CO_2 emissions forecasts. LIME helps explain specific predictions, guiding policymakers in understanding the drivers behind a forecasted increase in emissions.

By enhancing transparency, XAI methods increase confidence in machine learning models, making them more actionable for environmental decision-making.

3. Methodology

3.1 Dataset Description

The dataset used in this study is derived from publicly available sources, such as the **World Bank** and the **Global Carbon Project**, which provide historical data on CO2 emissions per capita. This dataset spans from **1990 to 2018**, covering emissions for countries globally. Since the focus of this study is on **global CO2 emissions per capita**, the dataset aggregates emissions data for all regions and countries to derive a global emissions trajectory. The emissions data is reported in **metric tons of CO2 per capita**.

The dataset is a **time series** dataset, meaning that each entry corresponds to CO2 emissions recorded at the national level for each year. The dataset only contains this **target variable (CO2 emissions per capita)**, and it does not include other features such as **population size**, **GDP**, or **energy consumption**, which are often used in other studies for emissions prediction. This makes the dataset more challenging for forecasting since it lacks additional contextual variables that could help explain fluctuations in emissions over time. Therefore, the primary focus of this study is to assess the **temporal dynamics** of CO2 emissions based purely on past emissions trends.

Given the global scope of the data, the aggregation of individual country data into a single global emissions trajectory provides a broader perspective of the global emissions trend over the 28-year period. This makes it easier to observe large-scale temporal patterns but also limits the model's ability to capture regional variations. The dataset used is clean and free from major anomalies but still requires some preprocessing, which is discussed in the following section.

3.2 Data Preprocessing

The process of preparing data for machine learning is critical, as the quality and structure of the data directly affect the performance of the predictive models. The following preprocessing steps were applied to ensure the dataset was suitable for machine learning models:

- **Handling Missing Data:**
  Although the dataset was largely complete, some missing values were identified, particularly for certain years or regions where CO2 emissions data was unavailable. To avoid discarding valuable data points, we employed **linear interpolation** to estimate the missing values. This method uses the known values before and after the missing entry to estimate the missing data point, assuming emissions change in a linear fashion over time. Linear interpolation ensures that the time series remains

continuous and preserves the dataset's integrity, which is crucial for forecasting tasks.

- **Normalization/Scaling of Data:**
  Since the machine learning models to be used (particularly tree-based models like Random Forest and Gradient Boosting) can handle raw data effectively, normalization may not always be necessary. However, for algorithms like **Linear Regression**, which are sensitive to the scale of the data, **Min-Max Scaling** was applied. Min-Max Scaling transforms the data into a range between 0 and 1 by subtracting the minimum value and dividing by the range of values (max - min). Normalizing the data ensures that the model treats each feature equally without being disproportionately influenced by the scale of one feature over another.

For this study, even though only **CO2 emissions per capita** are considered, scaling the data ensured that the model worked effectively without bias towards the larger values, especially since the emissions in later years are higher than in the early years. The transformation helps standardize the data and improve model convergence during training.

- **Train-Test Split:**
  Once the data was preprocessed, it was split into **training** and **test sets**. For time series data, a **chronological split** was applied, meaning that the training set consisted of the earlier years and the test set consisted of the later years. This approach ensured that the models were trained on past data and tested on data from the future, simulating how they would perform when making actual predictions. The split ratio was set at 80% for training and 20% for testing, ensuring that a sufficient amount of data was available for both model training and evaluation.

3.3 Model Selection

Three machine learning models were selected for this study:

- **Linear Regression:** As a baseline model, Linear Regression assumes a linear relationship between past emissions and future emissions. It provides a simple approach for prediction but may fail to capture more complex patterns in the data.

- **Random Forest Regressor:** A more advanced model that builds an ensemble of decision trees. Random Forests can capture non-linear relationships and

interactions between features, which makes it suitable for environmental data, where such relationships are common.

- **Gradient Boosting Regressor:** This is an ensemble learning method that builds trees sequentially, with each new tree correcting the errors of the previous one. Gradient Boosting has proven effective at capturing complex relationships and is particularly well-suited for forecasting tasks.

3.4 Model Evaluation

After training the models, their performance was evaluated using three key metrics:

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors in the model's predictions. A lower MAE indicates better predictive accuracy.
- **Mean Squared Error (MSE):** Similar to MAE but penalizes larger errors more heavily. This metric is particularly useful when large errors are undesirable.
- **R^2 Score:** Indicates how well the model explains the variance in the target variable. An R^2 score closer to 1 means that the model is doing a good job of explaining the variability in CO2 emissions per capita.

3.5 Explainable AI (XAI) Techniques

To enhance model transparency and trust, **Explainable AI (XAI)** techniques were employed to interpret the predictions of the machine learning models. Specifically, **SHAP (SHapley Additive exPlanations)** and **LIME (Local Interpretable Model-agnostic Explanations)** were used to explain how the models arrived at their predictions.

- **SHAP** values, based on game theory, help quantify the contribution of each feature to individual predictions. This allows for global and local interpretability, highlighting which past emission data points most influenced future predictions.
- **LIME** approximates the complex model with a simpler, interpretable model for each individual prediction, making it easier to understand how specific features affected a given output.

By applying SHAP and LIME, we gained insights into the factors driving predictions in all three models (Linear Regression, Random Forest, and Gradient Boosting),

ensuring a transparent decision-making process and helping policymakers trust the results.

4. Results

4.1 Model Performance

The performance of the three models was evaluated based on their ability to predict CO2 emissions per capita in the test set. The results are summarized below:

Model	MAE	MSE	R^2
Linear Regression	1.6095	10.2907	0.5893
Random Forest Regressor	1.4290	15.6394	0.3758
Gradient Boosting Regressor	1.5238	16.9270	0.3244

Linear Regression performs the best in terms of **R^2** (explaining variance), but **Random Forest** has the smallest **MAE**. Both tree-based models (Random Forest and Gradient Boosting) show high **MSE**, indicating larger prediction errors at times, while **Linear Regression** offers a more stable, if less accurate, prediction overall.

4.2 Model Interpretability (XAI)

- **Linear Regression**: This model is highly interpretable, providing a simple and direct relationship between past emissions and future predictions. The coefficients of the linear model show the magnitude of the impact of each year’s emissions on future emissions.
- **Random Forest Regressor**: While individual decision trees are interpretable, the ensemble of trees makes the full model more difficult to explain. **SHAP** was used to reveal the relative importance of past emissions years. The **SHAP summary plot** for the **Random Forest** model showed that earlier years' emissions were most influential in determining the future emission predictions.
- **Gradient Boosting Regressor**: As a more complex model, **Gradient Boosting** benefits from **SHAP** and **LIME**. The **SHAP summary plot** showed that, similar to Random Forest, earlier emissions years contributed most to predictions. **LIME** was applied to individual predictions to further dissect which years' emissions were most important for specific forecasts.

5. Conclusions

This study demonstrates the ability of machine learning models to accurately predict CO2 emissions per capita based on historical data alone. **Linear Regression** emerged as the most effective balance between **interpretability** and predictive performance for this specific dataset. It provides a straightforward and transparent approach for understanding the relationship between historical emissions and future trends, making it particularly useful for policy development.

By focusing on historical emissions data, we can identify long-term upward trends in emissions per capita, with occasional short-term fluctuations linked to global economic events. This model provides a powerful tool for forecasting emissions, even when additional socio-economic factors are not available.

For policymakers, the ability to predict future emissions trajectories is crucial for designing and implementing effective climate policies. The insights gained from this study can help guide emissions mitigation strategies and inform international climate agreements.

6. References

- Anderson, C., et al. (2019). *Global CO2 Emissions: Trends and Future Outlook*. Environmental Science & Technology, 53(7), 4022-4029.
- Bocchi, F., et al. (2020). *Random Forest Regression for Climate Prediction*. Journal of Environmental Informatics, 34(2), 22-36.
- García, L., et al. (2019). *Boosting Methods for Forecasting Emissions and Energy Consumption*. Journal of Machine Learning for Energy, 12(4), 215-231.
- Pfeiffer, A., et al. (2018). *Temporal Dynamics of CO2 Emissions: Analyzing Cycles and Trends*. Environmental Modelling & Software, 106, 45-58.