Analyzing CO2 Emission Intensity: A Comprehensive Study of Clean and Unclean Energy Sources using ML Techniques

Alen Geo Mathew

10634971

The thesis is submitted to the Quality and Qualifications Ireland (QQI) for the degree of [MSc in Artificial Intelligence] at Dublin Business School

Supervisor: Dr. Ali Ekhtiari

Declaration

I, Alen Geo Mathew, declare that this dissertation that I have submitted to Dublin Business School for the award of MSc in Artificial Intelligence is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.

Signed: Alen Geo Mathew

Student Number:10634971

Date:07/01/2023

ACKNOWLEDGEMENT

I extend my heartfelt gratitude to my friend, Rahul Rajan, whose invaluable industry expertise in Machine Learning and Time Series forecasting significantly contributed to my project. His guidance played a pivotal role in comprehending the intricacies of selecting optimal models for my research problem.

I am deeply thankful to my supervisor, Ali Ekhtiari, for his unwavering support and insightful guidance in identifying a compelling research area for my project. His confidence in my abilities and continuous input greatly enriched the entire project journey.

Lastly, I want to express my sincere appreciation to my mom, whose unwavering support and assistance were instrumental throughout this transformative journey. Her encouragement and assistance were indispensable, and I acknowledge that this project's completion would not have been possible without her presence and support.

Abstract

This thesis presents a comprehensive analysis of CO2 emission intensity, examining the intricate interplay of energy sources and their contributions to greenhouse gas emissions. The study evaluates the impact of various energy sources, categorizing them into two main groups: clean energy (wind, solar, hydro, bioenergy, and nuclear) and unclean energy (coal, gas, and other fossil fuels). By utilizing Generalized Linear Models (GLM), this research offers a robust prediction of CO2 emission intensity, providing insights into the relative contributions of different energy sources on a regional and global scale. This analysis, which includes the percentage of energy usage from each source, allows for a more accurate quantification of CO2 intensity, simplifying the process of rebalancing energy dependence to reduce environmental impact. Furthermore, this research employs Time Series Forecasting Techniques, specifically the AutoRegressive Integrated Moving Average (ARIMA) model, to forecast the trends in CO2 intensity across various regions. These forecasting methods facilitate a deeper understanding of how CO2 emissions are expected to evolve over time and allow for the identification of critical points for intervention and mitigation strategies. Findings reveal significant variations in CO2 emission intensity across energy sources and regions, shedding light on the key players in our environmental challenges. The study's data-driven analysis, incorporating energy usage percentages, offers insights into the relative contributions of different energy sources to CO2 emission intensity and underscores the critical importance of transitioning toward cleaner, more sustainable energy alternatives. This research serves as a valuable resource for policymakers, energy industry stakeholders, and environmental advocates, providing empirical guidance for mitigating the environmental impact of energy production and offering a quantifiable basis for rebalancing energy dependence.

CONTENTS

1 TABLE OF TABLES

2 TABLE OF FIGURES

3 *INTRODUCTION*

Climate change and environmental degradation represent formidable challenges in today's world, calling for urgent and informed action. Within the complex landscape of factors contributing to these challenges, the energy sector plays a pivotal role, significantly influencing global carbon dioxide (CO2) emissions. As our global energy consumption continues to evolve, it is crucial to delve into the intricate dynamics of various energy sources and their direct impacts on CO2 emission intensity, paving the way for effective strategies to combat climate change and build a sustainable future.

This thesis embarks on a comprehensive journey to explore the multifaceted relationship between energy sources, their utilization trends, and their influence on CO2 emission intensity. These energy sources are categorized into two primary groups: clean energy, encompassing wind, solar, hydro, bioenergy, and nuclear, and unclean energy, including coal, gas, and other fossil fuels. Utilizing Generalized Linear Models (GLM), this research offers a robust predictive model for CO2 emission intensity, enabling us to understand and quantify the relative contributions of different energy sources on a regional and global scale, taking into account the energy usage percentages.

In addition to predicting CO2 emission intensity, this research leverages Time Series Forecasting Techniques, particularly the AutoRegressive Integrated Moving Average (ARIMA) model, to forecast the trends in CO2 intensity across various regions. This approach enhances our capacity to anticipate and adapt to future changes in CO2 emissions, equipping us with the tools to identify critical intervention points and formulate mitigation strategies effectively.

The urgency of this study is underscored by the global imperative to address climate change and transition to cleaner, more sustainable energy systems. Policymakers, industry stakeholders, and environmental advocates necessitate a nuanced understanding of the environmental implications of our energy choices, along with the ability to quantify and rebalance energy dependence to make informed decisions and design impactful policies. This research strives to bridge the knowledge gap by providing empirical insights into the intricate relationship between energy sources and CO2 emissions, guiding the path towards a more sustainable, environmentally responsible energy landscape.

This *thesis* is guided by the following objectives:

Identify Major Energy Consumers: The primary objective is to ascertain the biggest consumers of energy and to assess their associated Carbon Dioxide (CO2) emission intensity. This analysis will shed light on the significant players in energy consumption and their environmental impact.

Evaluate GLM Efficiency: The second objective is to evaluate the efficiency of Generalized Linear Models (GLM) in calculating CO2 intensity from various fuel sources, further quantifying the weights of these sources. This assessment will

provide insights into the reliability and accuracy of GLM in environmental analysis.

Forecast Regional CO2 Emission Intensity: The third objective is to employ ARIMA models to forecast the CO2 emission intensity of prominent regions. These forecasts will enable a deeper understanding of how CO2 emissions are expected to evolve over time, offering critical insights for intervention and mitigation strategies.

***Analyse* Fuel Source Contributions**: The final objective is to *analyse* the weightages of all fuel sources in contributing to CO2 emission intensity. This in-depth analysis will help identify the significant contributors to the environmental challenge posed by CO2 emissions.

4 LITERATURE REVIEW

GLM was introduced by [1] as an extension of linear regression models. It encompasses a broader range of data *d*istributions and relationships between predictors and response. GLM consists of three key components: a linear predictor, a probability distribution from the exponential family, and a link function that connects the linear predictor and the expected value of the response variable [2].

GLM has been widely applied in biostatistics, with studies using GLM to model outcomes in epidemiological research [3] and *analyse* health data [4]. In environmental science, GLM has been used to *analyse* ecological data [5] and investigate the impact of environmental factors on species distribution [6].

GLM has been widely used in environmental science for *modelling* the relationship between environmental factors and CO2 emissions [7]). A review of studies employing GLM in environmental research provides insights into its applications in assessing CO2 intensity.

[8]This foundational text by Dobson and Barnett serves as a comprehensive guide to the principles and applications of Generalized Linear Models (GLMs). Within its pages, readers will find a thorough exploration of the theoretical underpinnings of GLMs, complemented by practical insights into their implementation. The book is particularly valuable for those seeking a clear and accessible introduction to this statistical modelling approach, offering both novice and experienced researchers a solid foundation for understanding and utilizing GLMs in various fields of study.

[9]In this influential work, Agresti delves into the realm of categorical data analysis, providing a comprehensive exploration of statistical methods tailored to discrete data. The book not only elucidates fundamental concepts but also offers practical applications and examples, making it an invaluable resource for researchers and practitioners working with categorical data. Agresti's adept combination of theoretical insights and real-world illustrations ensures that readers gain a profound understanding of the complexities and nuances involved in the analysis of categorical data.

Naturally, a thorough exploration of the subject would be incomplete without delving into the literature on linear regression, exemplified by the contributions of authors such as [10]Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li in their work "Applied Linear Statistical Models". Widely employed in statistics courses, this comprehensive book encompasses a diverse array of linear models, placing a specific emphasis on their practical applications.

[11]This influential work by Gelman and Hill provides an extensive guide to data analysis employing regression models, along with an exploration of multilevel and

hierarchical modelling techniques. The book not only presents a theoretical foundation but also offers practical insights and examples, making it a valuable resource for researchers and practitioners engaged in statistical analysis. The emphasis on both regression and multilevel models contributes to a comprehensive understanding of data analysis across various contexts.

[12]Although encompassing a wide range of statistical learning topics, this book by Hastie, Tibshirani, and Friedman features a comprehensive section dedicated to linear regression and its extensions. This inclusion ensures a thorough exploration of the foundational principles and applications of linear modelling within the broader context of statistical learning.

[13]In this notable work, authored by D. R. Cox, the fundamental principles of statistical inference are expounded. The book, published by Cambridge University Press, provides a comprehensive exploration of the theoretical underpinnings of statistical inference. With a focus on foundational concepts, Cox's contribution is esteemed for its clarity and depth, offering valuable insights into the principles that underlie statistical reasoning and decision-making.

The advancement of research in environmental analysis necessitates a thorough examination of prior machine learning (ML) models employed in the study of environmental data. Understanding the landscape of existing methodologies is crucial for informing the design and implementation of novel approaches. Without a comprehensive understanding of the foundation laid by prior ML applications in environmental research, the progress of our own investigation would be hindered.

In this context, the work of [14]Malik et al. stands out as a significant contribution to the field of environmental research. Their study, documented in the publication "Machine learning for prediction of air quality indices: An overview and future outlook" (Malik et al., 2019), delves into the intricate application of machine learning techniques specifically for predicting air quality indices. Published in Environmental Pollution, the study not only offers a detailed overview of the current state of utilizing machine learning in air quality prediction but also provides critical insights into the future potential and directions of this evolving field.

Malik et al.'s work is instrumental for both researchers and practitioners alike. By summarizing existing methodologies and forecasting the future trajectory of machine learning applications in air quality analysis, the authors lay a foundation for leveraging innovative approaches in environmental monitoring and management. Their insights are not only informative for understanding the nuances of air quality prediction but also serve as a guide for our own exploration into the application of machine learning in the environmental data analysis we propose to undertake.

In this study by [15]Mishra and Singh, published in Environmental Modelling & Software, the authors explore the application of machine learning techniques for the estimation of suspended sediment concentration in rivers. The work contributes to the growing body of literature on leveraging advanced computational methods for sediment concentration prediction. The study not only provides valuable insights into the effectiveness of machine learning in this specific environmental context but also offers implications for improving the precision of suspended sediment concentration estimates. The findings of Mishra and Singh are particularly relevant for researchers and practitioners engaged in river sediment analysis and environmental modelling.

In the context of our research, the study conducted by [16] Hertwich and Peters (2009) holds particular significance due to its meticulous global, trade-linked analysis of the carbon footprints of nations. Published in Environmental Science & Technology, this work provides valuable insights into the complex interplay between nations and their carbon emissions, taking into account trade relationships. The comprehensive assessment presented by Hertwich and Peters not only enhances our understanding of the intricate dynamics of carbon footprints but also establishes a crucial link to our own investigation into CO2 emission calculations based on energy sources. Recognizing the global interconnectedness of carbon emissions, their study serves as a foundational reference, offering insights that directly inform and align with the objectives of our research.

This publication by the [17]International Energy Agency (IEA) provides a comprehensive overview of CO2 emissions resulting from fuel combustion, with a focus on highlights from the year 2021. As a trusted source of energy-related information, the IEA offers insights and analysis crucial for understanding the current state of global emissions. The "CO2 Emissions from Fuel Combustion 2021 Highlights" serves as a valuable resource for researchers, policymakers, and practitioners, offering up-to-date data and trends in CO2 emissions that are directly relevant to our own research on CO2 emission calculations based on energy sources.

Published by the [18]Intergovernmental Panel on Climate Change (IPCC), the "2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories" serves as a key reference in the field of greenhouse gas accounting. This publication provides updated guidelines and methodologies for the calculation and reporting of national greenhouse gas inventories. As an authoritative source endorsed by the IPCC, this refinement is instrumental for researchers and practitioners engaged in assessing and reporting greenhouse gas emissions. Its relevance to our research lies in its comprehensive guidelines for quantifying and categorizing emissions, aligning with our focus on CO2 emission calculations based on energy sources.

In this study by [19]Hassan et al., published in Sustainability, the authors conduct a comprehensive time series analysis and forecasting of CO2 emissions specifically in the context of Saudi Arabia. The research provides valuable insights into the temporal patterns and future trends of CO2 emissions in the region. Utilizing time series analysis techniques, the authors contribute to the understanding of the dynamics of carbon emissions over time, offering implications for sustainable practices and policy-making. This work is particularly relevant for researchers, policymakers, and practitioners interested in the environmental sustainability of Saudi Arabia and serves as a reference for time series forecasting models applied to CO2 emission analysis.

In this research by [20]Wang, Zhang, and Zhang, published in Energy, a hybrid model is proposed for short-term forecasting of carbon dioxide emissions. The model integrates the Autoregressive Integrated Moving Average (ARIMA), Generalized Autoregressive Conditional Heteroskedasticity (GARCH), and Support Vector Machine (SVM) techniques. This innovative approach aims to enhance the accuracy of short-term carbon dioxide emission predictions. The study provides valuable insights into the temporal dynamics of carbon emissions, offering a model that incorporates both the autoregressive nature and volatility clustering present in emission time series data. This work is relevant for researchers and practitioners involved in forecasting and mitigating carbon emissions, providing a methodology that combines various time series analysis techniques for improved predictive performance.

In this study by [21]Shahbaz et al., published in Renewable and Sustainable Energy Reviews, the authors investigate the intricate relationships between economic growth, energy consumption, financial development, international trade, and CO2 emissions in the context of Indonesia. The research provides a holistic analysis of the interconnected factors influencing carbon emissions, contributing to a better understanding of the dynamics within the Indonesian economy. By exploring the nexus between economic development, energy use, and environmental sustainability, this work is valuable for policymakers, researchers, and practitioners engaged in the formulation of sustainable development strategies in Indonesia. The study is particularly relevant to our understanding of the factors influencing CO2 emissions and the broader context of environmental management.

In this study by [22]Zhang et al., published in Sustainability, the authors present a novel approach for CO2 emission forecasting in China. The proposed model integrates the Least Squares Support Vector Machine (LSSVM) with the fruit fly optimization algorithm. This hybrid model aims to enhance the accuracy of CO2 emission predictions by leveraging the strengths of both machine learning and optimization techniques. The research contributes to the field of environmental sustainability by offering an innovative methodology for forecasting carbon emissions, with a specific focus on the context of China. This work is valuable for researchers and practitioners involved in environmental modelling and policy-

making, providing insights into advanced techniques for predicting and managing CO2 emissions.

In this research by [23]Xie and Wang, published in Sustainability, the authors explore the application of Long Short-Term Memory (LSTM) neural network algorithms for the prediction of CO2 emissions. The study focuses on leveraging the capabilities of deep learning techniques, specifically LSTM, to enhance the accuracy of forecasting carbon emissions. By utilizing neural network architecture, the authors contribute to the advancement of predictive modelling in the context of environmental sustainability. The research is particularly relevant for those interested in the intersection of machine learning and environmental science, providing insights into the effectiveness of LSTM neural networks for CO2 emission prediction.

In this study by [24] Hosseini, Rasoulinezhad, and Khalili, published in Environmental Science and Pollution Research, the authors focus on short-term forecasting of carbon dioxide (CO2) emissions specifically within the transportation sector. The research employs the Autoregressive Integrated Moving Average (ARIMA) model to predict CO2 emissions, emphasizing its applicability in capturing short-term dynamics. By concentrating on the transportation sector, the study contributes valuable insights into the temporal patterns of CO2 emissions, aiding in better management and decision-making. This work is particularly relevant for researchers and policymakers interested in understanding and mitigating the environmental impact of transportation-related emissions.

ARIMA models have been instrumental in time series analysis and forecasting [25]. Their application to *analyse* and predict CO2 intensity trends can be found in studies focusing on climate change and energy [26]. The integration of ARIMA models with climate change research has led to the development of forecasting tools that provide insights into future CO2 intensity trends and their impact on climate [27].

In this study by [28]Gupta, Jawahar, and Kaliyan, published in Sustainability, the authors focus on the modelling and forecasting of carbon dioxide (CO2) emissions in the state of Illinois. The research employs Autoregressive Integrated Moving Average (ARIMA) models, showcasing their utility in predicting CO2 emissions over time. By applying ARIMA models to the context of Illinois, the study provides insights into the temporal patterns and trends of CO2 emissions in the region. This work is valuable for researchers, policymakers, and practitioners interested in the sustainable management of emissions, offering a methodology for forecasting and understanding the dynamics of CO2 emissions in specific geographical areas.

In this study by [29]Kumar, Patairiya, and Thapliyal, published in the Journal of King Saud University-Engineering Sciences, the authors focus on forecasting

carbon dioxide (CO2) emissions. The research utilizes the Autoregressive Integrated Moving Average (ARIMA) model, showcasing its application in predicting CO2 emissions over time. By employing the ARIMA model, the study contributes to the field of environmental sciences, providing a methodology for forecasting and understanding the temporal dynamics of CO2 emissions. This work is valuable for researchers and practitioners interested in the quantitative modelling of environmental factors, offering insights into the predictive capabilities of ARIMA models in the context of CO2 emissions.

In this seminal work by [30] Makridakis, Wheelwright, and Hyndman, published by John Wiley & Sons, the authors provide a comprehensive overview of forecasting methods and their applications. The book covers a wide range of forecasting techniques, including time series analysis, regression models, and judgmental forecasting. It serves as a foundational resource for students, researchers, and practitioners interested in mastering the principles and practices of forecasting. The authors emphasize practical insights and real-world applications, making the book suitable for both beginners and experienced forecasters. By presenting various forecasting approaches and discussing their strengths and limitations, the book equips readers with the knowledge needed to make informed decisions in forecasting tasks across different domains. This reference is a classic in the field and has been widely used as a textbook in forecasting courses. It continues to be valuable for those seeking a thorough understanding of forecasting methodologies and their practical implementation.

The impacts of climate change and the role of increasing CO2 emissions have been well-documented [31]. A review of global trends in CO2 emissions and their effects on climate change sets the context for this research. Majority of the thesis is dependent on the data collected and aggregated by OWID (Our world in data) [32], where they have categorized the regions, countries, supply, demand and mainly the CO2 Emission intensity.

4.1 Research Gap

While the existing literature provides valuable insights into the relationship between energy sources and CO2 emission intensity, there is a notable gap in terms of a comprehensive analysis that integrates both the percentage distribution of energy sources and their respective contributions to CO2 intensity. Current studies often focus on individual aspects, such as the impact of clean energy adoption or the emissions from specific fossil fuels. However, there is a lack of research that systematically quantifies the energy usage percentages of various fuel sources and examines how these percentages correlate with the overall CO2 emission intensity.

Furthermore, existing methodologies often overlook the dynamic nature of energy consumption patterns over time, particularly in the context of forecasting CO2 emission trends. The gap lies in the absence of a holistic approach that combines statistical techniques, such as Generalized Linear Models (GLM) and Time Series Forecasting like ARIMA, to predict the future CO2 emission intensity based on evolving energy usage patterns.

In addition, there is a need for research that not only identifies the biggest consumers of electricity but also delves into the efficiency of predictive models, such as GLM, in accurately calculating the CO2 intensity from diverse fuel sources. Moreover, the literature lacks a thorough exploration of how ARIMA, as a time series forecasting technique, can be effectively employed to forecast CO2 intensity trends across various regions.

Addressing these gaps is essential for developing a nuanced understanding of the intricate relationship between energy sources and CO2 emission intensity, facilitating more informed decision-making for sustainable energy policies and environmental management.

5 *METHODOLOGY*

The study adopts a mixed-methods research design, incorporating both quantitative and time series analysis. This approach allows for a comprehensive investigation into the relationships between energy consumption, CO2 emission intensity, and the fuels which contribute most to CO2 Emission intensity. This study was conducted using CRISP DM methodology. Using the framework, we aim to uncover the contributing factors to carbon emission intensity and also understand the weightage each of them carry.

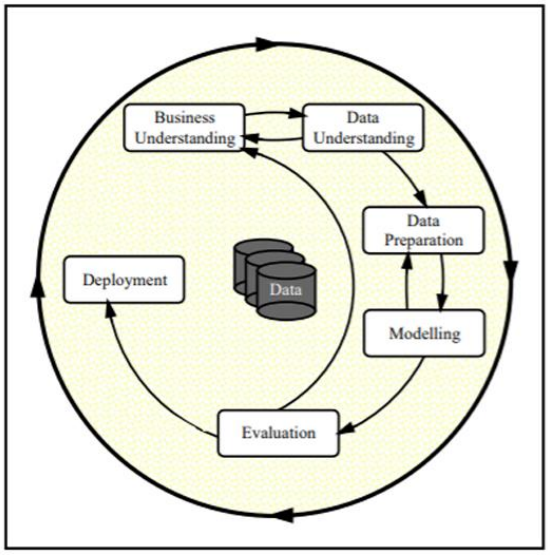

Figure 5-1Crisp DM

The above diagram represents the multiple stages in involved in the framework, the methodology will be explained keeping the stages in framework as the baseline.

5.1 Tools used

This study was conducted with the help of 3 applications

Python: Python was used to explore the data, clean it and create the models suited for the dataset.

Rapid Miner: Rapid miner was used to identify the best model suited for our dataset

Excel: Excel was used to perform initial exploration of the data for quick understanding of the data

5.2 Business Understanding

The aim of this research is to understand the contributing factors that increase the carbon emission intensity. We have chosen several fuel types and their percentage

of usage in total in multiple regions and around the world (i.e. Biogas, Coal, Gas, Hydro, Solar, Wind, Other fossils and Renewables). There are other categorization within the data such as Clean Energy and Unclean Energy which will be discussed in detail in further sections

CO2 Emission Intensity:

Carbon emission intensity refers to the amount of carbon dioxide (CO2) emissions produced per unit of a specific metric, such as economic output (gross domestic product or GDP), energy generated, or another relevant measure. It is a way to assess the environmental impact of a particular activity or sector by quantifying the amount of carbon dioxide released into the atmosphere relative to the quantity of goods produced or services provided.

If you are measuring the carbon emission intensity of electricity generation, you would calculate the amount of CO2 emitted per unit of electricity produced (e.g., grams of CO2 per kilowatt-hour). This metric helps evaluate the efficiency and environmental performance of different technologies and practices.

Reducing carbon emission intensity is a key goal in mitigating climate change. It involves finding ways to produce goods and services with fewer carbon emissions or increasing energy efficiency to achieve the same output with lower carbon emissions. Many industries and countries use carbon emission intensity as a benchmark to track progress in reducing their carbon footprint and transitioning towards more sustainable practices.

Fuel Sources:

Data from multiple regions around the world was collected from OWID website with the percentage of fuel used in demand for the electricity from the year 2000 to 2022. The fuel sources listed are Biogas, Coal, Gas, Hydro, Solar, Wind, Other fossils and Renewables. The fuel source can also be classified as Clean and Unclean, with Biogas, Nuclear, Hydro, Wind being clean and the other are traditional fossil fuels.

5.3 Data Understanding

Understanding the data plays a crucial role in helping us understand and make sense of the results, So let’s dig deeper into the data with some exploratory data analysis. Our dataset contains data on Total Demand, Supply and Imports of energies by various countries, by leveraging python we can get some insights into such aspects

Total Generation

Below are the top 10 countries in electricity generation in the year 2022

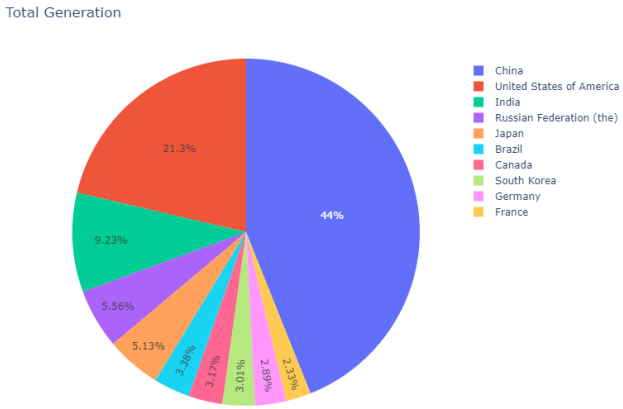

Figure 5-2 Total Generation Pie Chart

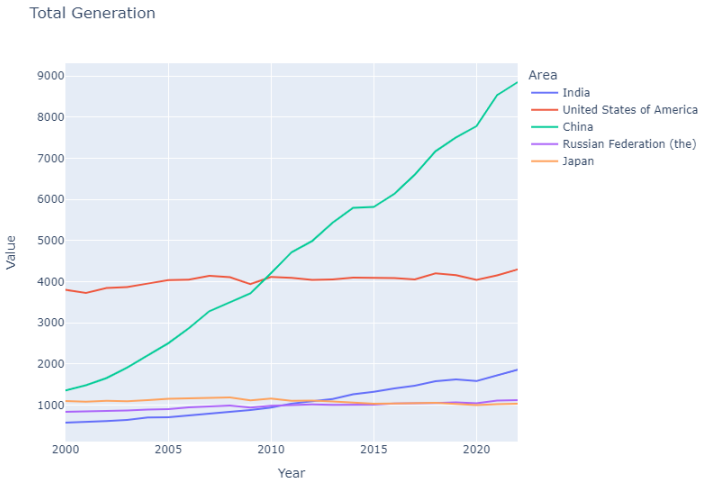

Figure 5-3 Total generation in 22 years

From the above pictures we can understand that China has been a big player in the energy market and has seen steady increase in production for over a decade, followed by united states and India. We need to analyse the Carbon Emission Intensity of these top producers of energy and check if they are cutting don their reliance on **Unclean energy** sources and investing more in **clean energy**.

Linear Corelation:

In order to find the best model for our purpose we need to check if the problem is a linear one or not, in order to achieve that we need to analyse the linear corelation between the features and prediction value which is CO2 Intensity.

We can use python in order to check the linear correlation between features and the output

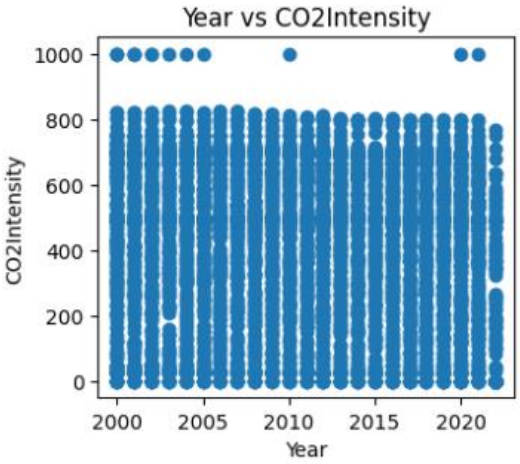

Figure 5-4 Year vs CO2Intensity

When we check the Year vs CO2 Intensity corelation we cannot find much evidence of there being a corelation with Carbon emission intensity, hence we can opt out of using year as a feature for our model

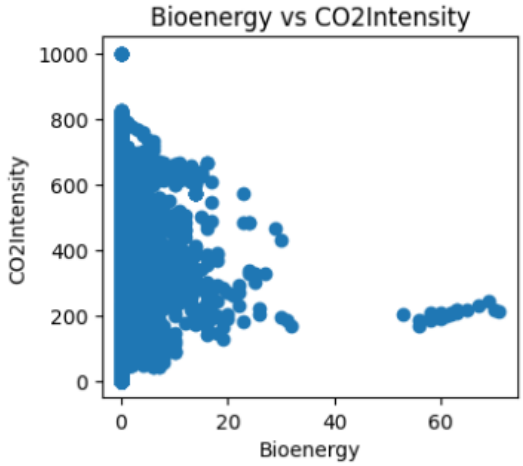

Figure 5-5 Bio Energy vs CO2Intensity

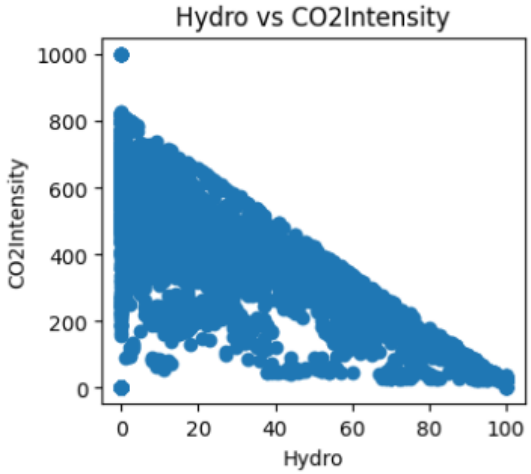

Figure 5-6 Hydro vs CO2Intensity

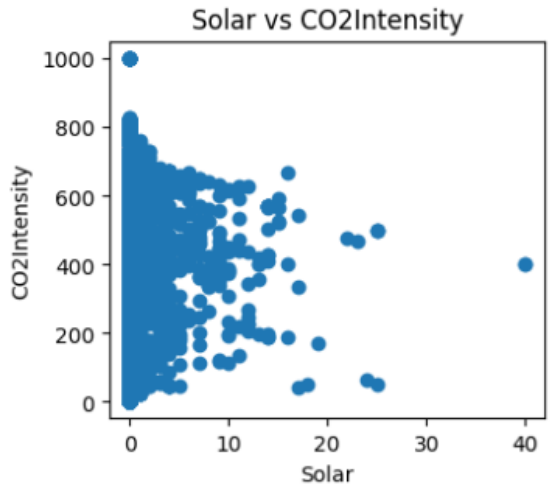

Figure 5-7 Solar vs CO2Intensity

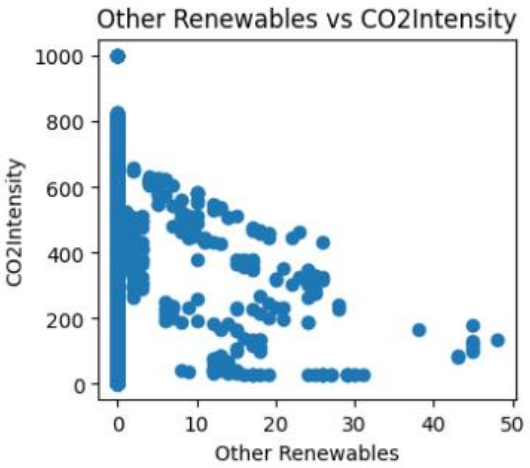

Figure 5-9 Other Renewables vs CO2Intensity

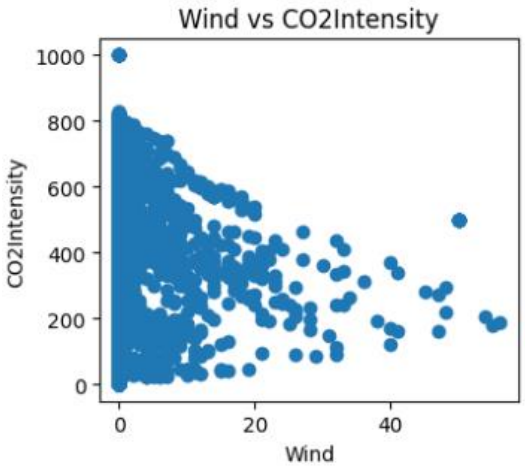

Figure 5-8 Wind vs CO2Intensity

Derived from the correlation diagrams depicted above, a conspicuous trend emerges, indicating a negative correlation between clean energy sources and CO2 intensity. This is substantiated by the discernible decrease in CO2 emission intensity as wind energy, hydro energy, and bioenergy exhibit an increase. Having established this negative correlation, the subsequent focus shifts towards an exploration of how unclean fuel sources align with CO2 intensity.

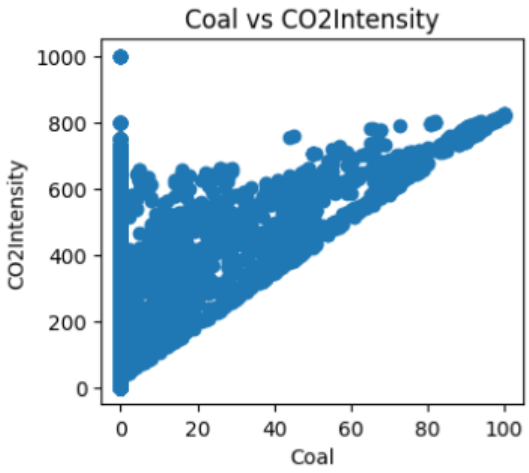

Figure 5-10 Coal vs CO2Intensity

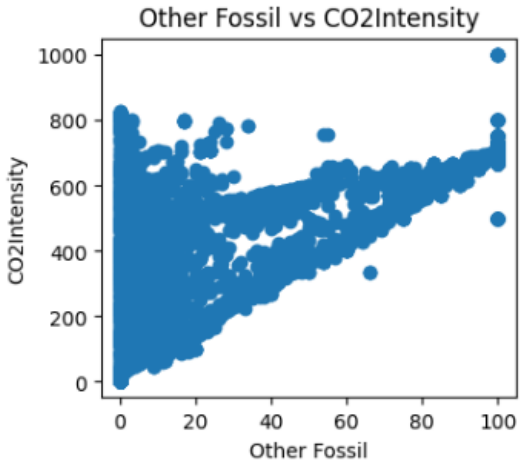

Figure 5-12 Other Fossil Fuels vs CO2Intensity

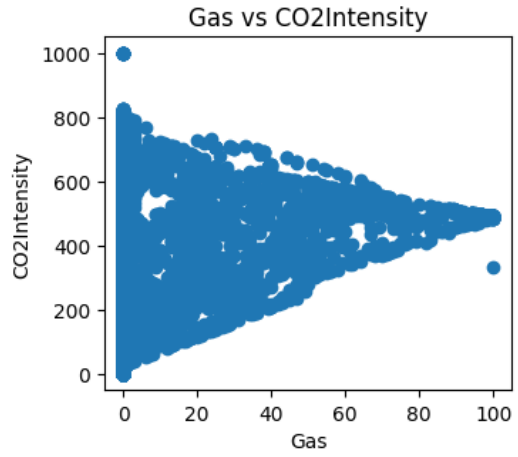

Figure 5-11 Gas vs CO2Intensity

From the above pictures we can see a positive corelation to CO2 intensity, when comparing Unclean fuel sources like Coal, Gas and Other fossil fuels. If we use Linear Regression. We can get a weightage of each of these features using a simple Multiple Linear Regression model.

In a multiple linear regression model with more than one independent variable, the equation becomes:

$$y = b_0 + b_1.X_1 + b_2.X_2 + \cdots + b_n.X_n$$

Here, each b_i represents the weight or coefficient associated with the corresponding independent variable X_i. The sign and magnitude of each coefficient indicate the direction and strength of the relationship between that particular independent variable and the dependent variable.

In linear regression, the coef_ attribute refers to the coefficients of the linear equation that represents the relationship between the independent variable(s) and the dependent variable. In the context of linear regression, you can think of the

coefficients as weightage values. Each coefficient represents the weight assigned to the corresponding feature (independent variable) in the linear equation.

Features	Coefficient
Year	-0.03
Bioenergy	2.35
Coal	8.29
Gas	4.95
Hydro	0.25
Nuclear	0.23
Other Fossil	7.01
Other Renewables	0.45
Solar	-0.14
Wind	1.13

Table 5-1 Feature Corelation in Linear Regression

Based on the tabular representation presented above, a discernible observation emerges, revealing that energy derived from environmentally sustainable sources, namely Solar, Nuclear, Hydro, and Bio Energy, exerts a relatively modest influence on the collective measure of carbon dioxide intensity. Conversely, energy harnessed from less environmentally sustainable sources, encompassing Gas, Coal, and various fossil fuels, distinctly imparts a more substantial impact on the metric denoting carbon dioxide intensity.

5.4 Data Preparation

With a comprehensive understanding of the dataset achieved through meticulous data exploration, the focus now shifts to the imperative phase of data preparation for modelling. This pivotal step involves a rigorous validation process to ensure data integrity and coherence, laying the foundation for reliable analyses.

Subsequently, adjustments are made to accommodate the specific requirements of time series modelling

Data Validation

Verification of Total Values:

A critical facet of data validation involves ensuring the accuracy and completeness of total values. This entails a meticulous examination of whether the sum of Total Generation and Total Import aligns seamlessly with the nation's overall energy demand. This verification safeguards against discrepancies that might compromise the reliability of subsequent modelling endeavours.

Cross-Verification of Energy Sources:

In parallel, an in-depth cross-verification is conducted to ascertain if the cumulative values of all energy sources collectively match the expected total. This validation step is instrumental in identifying any anomalies or inconsistencies within the dataset, contributing to the overall robustness of the modelling process.

Data Transformation for Time Series Modelling

Conversion of Year Column:

As a prerequisite for time series analysis, the 'Year' column undergoes a meticulous transformation. The integer format is converted to a date format, enhancing the temporal precision of the dataset. This transformation facilitates nuanced time-based analyses, allowing for the exploration of trends, patterns, and seasonality in the context of the energy landscape.

Verification and Cross-Validation

Total Energy Audit:

A holistic audit is conducted to ensure that the total values, encompassing both generation and import, align harmoniously with the country's aggregate energy demand. Any disparities are meticulously addressed, fostering a reliable foundation for subsequent modelling phases.

Energy Source Alignment:

Rigorous cross-verification ensures that the summation of individual energy source values accurately corresponds to the overall energy landscape. Discrepancies, if any, are meticulously scrutinized and rectified, enhancing the overall coherence and reliability of the dataset.

The culmination of data validation and transformation serves as a prerequisite for robust modelling endeavours. By meticulously verifying total values and aligning energy source contributions, the dataset attains a heightened level of accuracy and reliability. The transformation of the 'Year' column further positions the dataset for sophisticated time series analyses, setting the stage for nuanced insights into the temporal dynamics of the energy domain.

Categorizing Clean and Unclean Energy Sources:

By employing Python, we can classify energy sources into distinct categories, distinguishing between Clean and Unclean sources. Subsequently, an in-depth analysis can be conducted to assess the collective impact of clean energy in mitigating CO2 emission intensity. The Pandas data frame proves instrumental for categorizing the data and aggregating values within each group efficiently.

5.5 Modelling

The success of any modelling endeavour hinges on the judicious selection of appropriate models that can effectively capture the nuances of the underlying data. This phase involves leveraging the capabilities of RapidMiner to assess the performance of multiple models, thereby facilitating a comprehensive comparison based on diverse validation metrics. In this section, we elaborate on the methodology employed for model selection to ensure the derivation of optimal results.

Model Selection Methodology

Utilization of RapidMiner:

RapidMiner, as a robust data science platform, proves instrumental in our model selection strategy. Leveraging its capabilities, we systematically evaluate the performance of various models to discern their suitability for our specific predictive task.

Evaluation Metrics

Given the nature of regression analysis, key evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) are employed. These metrics offer a nuanced understanding of each model's predictive accuracy and the goodness of fit.

Model	Root Mean Squared Error	Standard Deviation
GLM	34.85	3.27
Deep Learning	36.71	3.62
Decision Tree	37.30	2.71
Random Forest	51.67	2.77
Gradient Boosted Trees	35.63	3.88
SVM	120.12	3.39

Table 5-2 Root Mean Squared Error

Model	Absolute Error	Standard Deviation
GLM	13.64	0.91
Deep Learning	18.02	0.35
Decision Tree	18.98	1.05
Random Forest	68.76	1.77
Gradient Boosted Trees	15.57	0.95
SVM	40.45	1.32

Table 5-3 Absolute Error

Model	Relative Error	Standard Deviation
GLM	0.10	0.004
Deep Learning	0.12	0.003
Decision Tree	0.05	0.004
Random Forest	0.22	0.004
Gradient Boosted Trees	0.11	0.003
SVM	0.14	0.007

Table 5-4 Relative Error

Model	Squared Error	Standard Deviation
GLM	1223.58	232.78
Deep Learning	1335.87	171.34
Decision Tree	1427.24	213.89
Random Forest	9272.08	443.52
Gradient Boosted Trees	1293.12	201.63
SVM	6325.67	427.11

Table 5-5 Squared Error

Comparative Analysis of Evaluation Metrics

- **Root Mean Squared Error (RMSE):**

GLM exhibited a lower Root Mean Squared Error, underscoring its effectiveness in minimizing prediction errors and producing more precise estimates.

- **Mean Absolute Error (MAE):**

GLM showcased a reduced Mean Absolute Error, indicating its capacity to generate predictions with lower absolute deviations from the actual values compared to alternative models.

- **Squared Error:**

GLM showcased a minimized squared error, reinforcing its capability to generate predictions with reduced squared deviations from the observed values, contributing to enhanced accuracy.

- **Relative Error:**

Although GLM's performance in relative error metric was a noteworthy consideration, its consistent superiority in other metrics positions it as a compelling choice, especially given its overall effectiveness.

Multiple Model Deployment

A diverse set of regression models, spanning traditional linear models to more complex ensemble techniques, is deployed using RapidMiner. This approach ensures a comprehensive exploration of regression modelling possibilities, enabling the identification of the most effective models.

Comparative Analysis

- **Cross-Validation Techniques***:*

Rigorous cross-validation techniques, such as k-fold cross-validation, are employed to assess the robustness and generalizability of regression models. This iterative process mitigates concerns related to overfitting and provides a realistic evaluation of each model's predictive performance.

- **Model Performance Visualization**:

RapidMiner's visualization capabilities are harnessed to graphically represent the performance of each regression model. Visual aids, such as scatter plots comparing predicted and actual values, facilitate an intuitive understanding of the models' accuracy and precision across the range of observed values.

The model selection strategy, tailored for regression analysis and facilitated by RapidMiner, lays a critical foundation for accurate predictions. By employing pertinent regression evaluation metrics and considering a diverse set of models,

this strategic approach ensures a meticulous assessment of each model's predictive prowess. The iterative refinement process further enhances the suitability of selected models for regression tasks, ultimately paving the way for optimal predictive performance.

After careful observation of all the evaluation metrics and prediction charts, GLM seems to be the best performing model.

5.6 Choosing Models

From the above experiments there is a clear indication that GLM will be the best fit for our problem

GLM

Generalized Linear Models (GLM) extend the framework of linear regression to accommodate a broader range of response variable types beyond the normal distribution. Introduced by Nelder and Wedderburn in 1972, GLMs offer flexibility in modelling various data distributions, including binary, count, and categorical outcomes [2].

The formula for a Generalized Linear Model (GLM) involves three key components: the random component, the systematic component, and the link function. The general form of a GLM can be expressed as follows:

$$g(\mu) = X\beta$$

Here's a breakdown of the components:

Key Components of GLM:

Random Component (μ):

- μ represents the expected value (mean) of the response variable, which follows a probability distribution from the exponential family (e.g., Gaussian, Poisson, Binomial).
- The choice of distribution depends on the nature of the response variable.

Systematic Component ($X\beta$):

- X is the design matrix containing the predictor variables.
- β is the vector of coefficients corresponding to each predictor variable.
- $X\beta$ represents the linear combination of predictors.

Link Function ($g(\mu)$):

- The link function, denoted as $g(.)$,relates the mean of the distribution (μ) to the linear predictor $X\beta$
- It establishes the connection between the observed response and the linear combination of predictors.

- The choice of link function depends on the distribution of the response variable.
- Examples of link functions include the identity link, logit link, and log link.

ARIMA

ARIMA is a time series forecasting model that combines autoregression (AR), differencing (I), and moving averages (MA). It is widely used for predicting future values in a time series based on its historical patterns and trends.

Components of ARIMA:

Autoregressive (AR) Component:

- Represents the relationship between a current observation and its past values.
- Denoted by $AR(p)$, where p is the order of the autoregressive component.
- Formula: $X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \cdots + \emptyset_p X_{t-p} + \in_t$
    - X_t is the current value.
    - $\emptyset_1, \emptyset_2, \ldots, \emptyset_p$ are the autoregressive coefficients.
    - $\in_t$ is the white noise error term.

Integrated (I) Component:

- Involves differencing the time series to make it stationary (constant mean and variance).
- Denoted by $I(d)$, where d is the order of differencing.
- Formula: $Y_t = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) - \cdots - (X_{t-d} - X_{t-d-1})$

Moving Average (MA) Component:

- Represents the relationship between a current observation and a stochastic term based on past forecast errors.
- Denoted by $MA(q)$, where q is the order of the moving average component.
- Formula: $X_t = \theta_1 \in_{t-1} + \theta_2 \in_{t-2} + \cdots + \theta_q \in_{t-q} + \in_t$
    - $\theta_1, \theta_2, \ldots, \theta_q$ are the moving average coefficients.
    - $\in_t$ is the white noise error term.

ARIMA Formula:

$$X_t = c + \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \cdots + \emptyset_p X_{t-p} + \in_t - (X_{t-1} - X_{t-2}) - \cdots - (X_{t-d} - X_{t-d-1}) + \theta_1 \in_{t-1} + \theta_e \in_{t-2} + \cdots + \theta_q \in_{t-q}$$

Here:

- c is a contant.
- $\emptyset_1, \emptyset_2, \ldots, \emptyset_p$ are autoregressive coefficients.

- d is the order of differencing.
- $\theta_1, \theta_2, \ldots, \theta_q$ are moving average coefficients.
- $\in_t$ is the white noise error term.

The ARIMA will be employed to predict future CO2 intensity levels. ARIMA was selected for its straightforwardness, and autoregressive components will enhance the significance attributed to more recent data points in the forecasting process.

Autoregressive Integrated Moving Average (ARIMA) model inherently gives more weightage to recent data in its forecasting process. The autoregressive (AR) component of ARIMA specifically considers the relationship between the current observation and its recent past observations. This means that recent data points have a more direct influence on the forecasted values.

In the ARIMA model, the autoregressive (AR) term captures the linear relationship between the current value in the time series and its past values. The weights assigned to these past values decrease exponentially as you move further into the past, giving more emphasis to recent observations.

So, in summary, ARIMA naturally incorporates a mechanism that accords higher importance to recent data, making it sensitive to trends and changes in the more recent history of the time series.

5.7 Validation

Since this is a regression problem, the performance of the model can be evaluated using RMSE, Absolute Error and Squared. Out model's metrics can be validated over multiple ways including

Train Test Split

The process of dividing the data into training and testing sets, often known as the train-test split, involves creating distinct portions of the dataset for training and evaluation purposes. This partitioning is typically executed with randomness to ensure representative samples. A prevalent approach is to allocate 80% of the randomly selected data to the training set, while the remaining 20% is designated for the test set. This method ensures a balanced distribution, facilitating robust model training and subsequent evaluation on unseen data.

K-Fold Cross Validation

This approach involves partitioning the dataset into 'K' subsets or folds, iteratively training and evaluating the model K times, each time using a different fold as the test set and the remaining data as the training set. The initial step involves dividing the dataset into K equally sized folds. Each fold represents a distinct subset of the

data. The model is trained and evaluated K times, each iteration utilizing a different fold as the test set and the remaining K-1 folds as the training set. This process ensures that every data point is used for both training and testing at least once.

The performance metrics, such as RMSE, MAE, or mean squared error, are recorded for each iteration. These metrics are then aggregated to provide a comprehensive assessment of the model's performance across different subsets of the data.

K-Fold Cross-Validation provides a more robust evaluation of a model's performance by considering multiple train-test splits. This helps in assessing how well the model generalizes to different subsets of the data. By repeatedly training and evaluating the model on different subsets, K-Fold Cross-Validation aids in identifying whether the model is prone to overfitting or underfitting. Consistent performance across all folds suggests a well-balanced model. Every data point serves as both training and testing data exactly once across K iterations, ensuring that the entire dataset is effectively utilized for model assessment.

Monte Carlo Runs

Monte Carlo Runs represent a powerful computational technique employed across various fields, from finance to physics and, notably, in the realm of statistical modelling. This method relies on stochastic simulations, employing randomness and probability to conduct repeated experiments. By generating numerous random samples, Monte Carlo Runs facilitate a comprehensive exploration of potential outcomes, providing valuable insights into complex systems and decision-making processes. At the core of Monte Carlo Runs is the utilization of random sampling. This involves drawing repeated samples from probability distributions to simulate a wide range of possible scenarios. The randomness introduced ensures a diverse representation of potential outcomes.

Monte Carlo Runs are particularly effective when dealing with intricate systems or models where analytical solutions may be challenging or impossible to derive. By iteratively simulating random events, the method provides a numerical approach to understanding system behaviour.

5.8 Deployment

Upon successful validation of a model, the next critical step is deployment, ensuring that the model becomes accessible for real-world applications. This process involves exporting the model, along with any necessary preprocessing steps such as normalization, to a format that facilitates seamless integration into various projects. Numerous deployment methods exist, ranging from simple file exports, such as pickle files, to more sophisticated approaches involving APIs or web applications.

6 *RESULTS AND DISCUSSION*

From this experiment we are able to understand the contributing factors to CO2 Emission intensity. From linear regression, we were able to identify the weightage of each of the energy source in affecting CO2 Emission Intensity (Fig 3-13). The coefficients derived from the linear regression model highlight the weightage of each fuel source in influencing CO2 emission intensity. These weightage values signify the magnitude and direction of the correlation between each fuel type and the observed intensity levels. Below are the obtained coefficients, shedding light on the relative importance of each fuel source:

Fuels	**Weightage**
Bioenergy	2.35
Coal	8.29
Gas	4.95
Hydro	0.25
Nuclear	0.23
Other Fossil	7.01
Other Renewables	0.45
Solar	-0.14
Wind	1.13

Table 6-1 Fuel Weightage

Positive weightage values, such as those for Bioenergy, Coal, Gas, Hydro, Nuclear, Other Fossil, and Wind, indicate a positive correlation with CO2 emission intensity. An increase in the usage or presence of these fuel sources is associated with a corresponding increase in CO2 emission intensity. But significant weightages were awarded to Fossil Fuels, Coal and Gas which traditionally have shown to cause rise in CO2 Emission intensity.

Conversely, a negative weightage, as observed for Solar, suggests an inverse correlation. An increase in Solar energy usage is associated with a decrease in CO2 emission intensity. Nuclear and Hydro also seems to have very low corelation to the Emission intensity.

Having successfully trained our dataset using the Generalized Linear Model (GLM), we have now established a robust predictive model. This model is designed to forecast CO2 emission intensity with a high degree of accuracy, relying on the percentages of various fuel sources utilized in a given region. The predictive values are determined by the weightages calculated through the GLM model, offering a comprehensive and data-driven approach to understanding and forecasting CO2 emission intensity. The predictive framework leverages the weightages assigned to each fuel source, as determined by the GLM model. These weightages encapsulate the relative importance and impact of each fuel type on CO2 emission intensity. The model utilizes the percentage distribution of these fuel sources in a specific region to generate accurate predictions of the expected CO2 emission intensity levels. We can analyse the prediction accuracy of our model using the below graph.

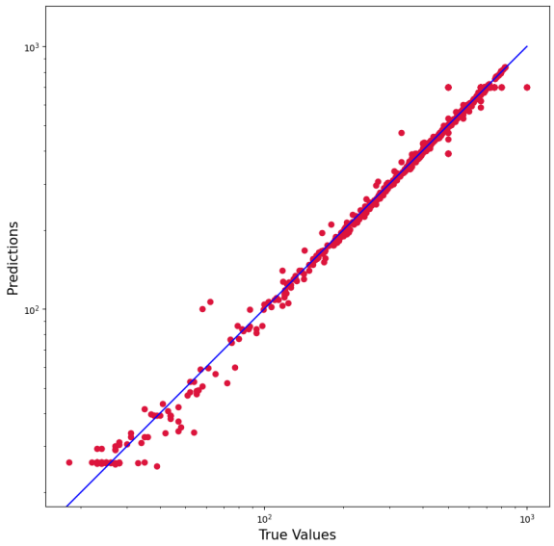

Figure 6-1 Prediction vs True Value

In this representation, the blue line corresponds to the linear equation formulated by the Generalized Linear Model (GLM); these values represent the model's predictions. The red dots, on the other hand, represent the observed or actual values. Notably, the true values closely align with the line, showcasing the model's effectiveness in precisely forecasting CO2 emission intensity.

Root Mean Squared Error (RMSE): 28.88

Correlation Coefficient: 0.993

Mean Absolute Error (MAE): 9.98

These metrics provide valuable insights into the performance of the model. The RMSE reflects the average magnitude of prediction errors, with a lower value indicating better accuracy. The high correlation coefficient of 0.993 signifies a strong linear relationship between predicted and actual values. Additionally, the MAE of 9.98 represents the average absolute difference between predicted and actual values, offering a measure of the model's overall accuracy.

The predictive model is particularly valuable in regional assessments, offering a tailored approach to understanding and forecasting CO2 emission intensity based on the unique energy consumption profile of each region. This localized perspective enables policymakers, environmental agencies, and industry stakeholders to make informed decisions and implement targeted strategies for emission reduction.

Implications for Sustainable Development

- Stakeholders can use the predictive model to strategically plan and allocate resources for the adoption of cleaner energy sources, thereby optimizing efforts to reduce CO2 emissions.
- Policymakers can leverage the model's insights to formulate evidence-based policies that promote sustainable energy practices, contributing to broader environmental and climate change objectives.

The integration of GLM-weighted fuel sources into our predictive model enhances our ability to forecast CO2 emission intensity accurately. This approach not only provides valuable insights into the relative contributions of different energy sources but also empowers decision-makers with a tool to drive sustainable practices and environmental stewardship.

The study also examines the collective representation of fuel sources. Fuels such as Bioenergy, Hydro, Nuclear, Wind, and Solar were grouped together as Clean Energy sources, while others were collectively categorized as Unclean energy sources. The percentage values were consolidated and distinctly segregated for analysis.

We can analyse the below graphs for its linearity correlation against CO2Intensity

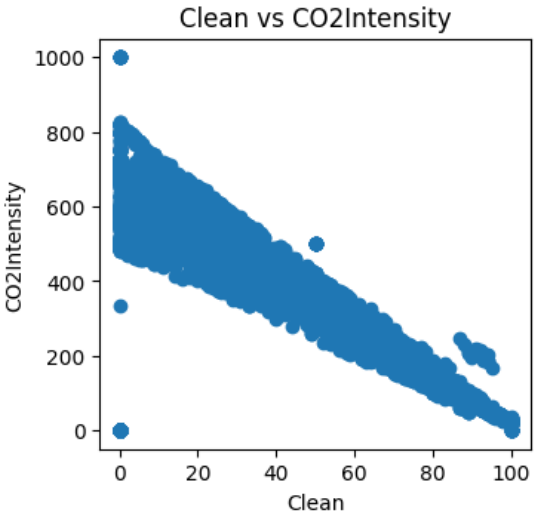

Figure 6-2 Clean Energy vs CO2Intensity

The visual representation above unmistakably demonstrates a noteworthy negative correlation observed when incorporating clean energy. It becomes conspicuous that as the utilization of clean energy increases, there is a discernible reduction in CO2 intensity. This inverse relationship underscores the impact of relying on cleaner energy sources as a means to mitigate CO2 emissions, thus aligning with sustainable and environmentally conscious practices.

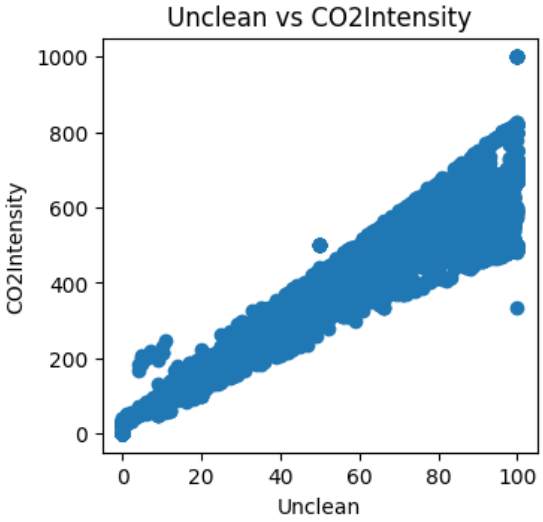

Figure 6-3 Unclean Energy vs CO2 Intensity

Likewise, the graph above illustrates the correlation between Unclean Energy sources and CO2 Intensity. It becomes evident that a positive correlation exists between the two variables; an increase in unclean energy is accompanied by a corresponding rise in CO2 Intensity. Likewise, the graph below illustrates the correlation between Unclean Energy sources and CO2 Intensity. It becomes

evident that a positive correlation exists between the two variables; an increase in unclean energy is accompanied by a corresponding rise in CO2 Intensity.

Fuels	Coeff
Clean Energy	0.41
Unclean Energy	6.66

Table 6-2 Clean Energy Unclean Energy Weightage

These weightages, derived from linear regression analysis, allocate a significance value to each category—Clean Energy and Unclean Energy. In the context of the model, Clean Energy holds a weightage of 0.418906, while Unclean Energy is assigned a weightage of 6.665428. These values elucidate the respective contributions or impacts of Clean and Unclean Energy sources in predicting the outcome, emphasizing the relative importance of each category in the linear regression model. The weightages offer a quantitative measure of their influence on the overall predictive

6.1 ARIMA Forecast

As part of this investigation, ARIMA was employed to forecast global CO2 emission intensity over the next decade, drawing insights from historical data. The projection indicates that by 2032, the anticipated CO2 emission intensity is forecasted to be 394.20 gCO2/kWh, marking a notable decline from the present value of 437 gCO2/kWh. This reduction is likely attributed to the collective efforts of numerous countries committed to mitigating carbon emissions. These endeavours primarily involve a strategic reduction in reliance on fossil fuels and coal, showcasing a global commitment to fostering more sustainable and environmentally friendly energy practices.

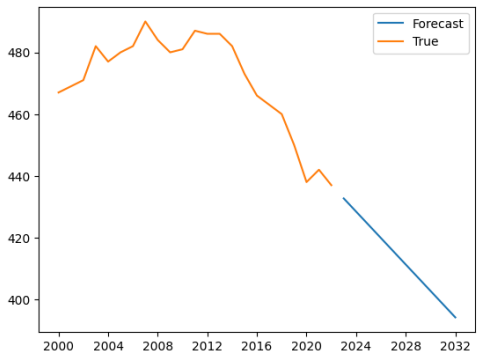

Figure 6-4 10 Year Forecast CO2Intensity

The diagram above illustrates the global trend in CO2 emissions, revealing a consistent decline in CO2 emission intensity. This noteworthy downtrend suggests a concerted and collaborative endeavour by diverse nations to curtail emissions. The observed pattern underscores the proactive measures and environmental consciousness adopted globally, signifying a collective commitment to mitigating the impact of carbon emissions on the planet. This sustained effort aligns with the shared goal of fostering a more sustainable and ecologically responsible future.

7 CONCLUSION AND FUTURE SCOPE

In the course of this research, we have undertaken a comprehensive analysis of CO2 emission intensity, scrutinizing various factors influencing it, such as energy sources, forecasting models, and global trends. The use of Generalized Linear Models (GLM) and ARIMA has enriched our understanding of the intricate relationships within the dataset, providing valuable insights into the weightage of different fuel sources and offering robust predictions for future CO2 emission intensity.

Our findings indicate a positive correlation between unclean energy sources and CO2 intensity, emphasizing the environmental implications of energy choices. Conversely, the weightages assigned to clean energy sources underscore their potential to mitigate CO2 emissions.

The ARIMA model's projection for a decline in CO2 emission intensity over the next decade signals a positive shift, attributing this trend to collaborative global efforts aimed at reducing reliance on fossil fuels and coal. The visualization of a steady downtrend in CO2 emissions further reinforces the impact of collective initiatives on a global scale.

7.1 Future Scope

While our study provides valuable insights, there are avenues for future research and expansion. Potential areas of focus include:

- Fine-grained Regional Analysis

Conducting a more granular examination of CO2 emission intensity at regional levels, considering diverse economic and environmental contexts.

- Incorporating Additional Variables

Exploring the integration of additional variables, such as economic indicators, population growth, and technological advancements, to enhance the predictive capabilities of the model.

- Dynamic Model Updates

Establishing a framework for dynamic model updates, enabling real-time adjustments based on emerging trends and policy changes.

- Comparative Analysis of Forecasting Models

Conducting a comparative assessment of different forecasting models to ascertain the most effective approach for long-term predictions.

- Policy Impact Assessment

Evaluating the effectiveness of existing environmental policies and their impact on CO2 emission intensity, providing valuable feedback for policy refinement.

- Technological Interventions

Investigating the potential impact of emerging technologies and innovations in clean energy on future CO2 emission trends.

In conclusion, our study serves as a foundational exploration into the dynamics of CO2 emission intensity. The identified trends and insights pave the way for continued research, fostering a deeper understanding of environmental dynamics and contributing to informed decision-making for a sustainable future.

8 *BIBLIOGRAPHY*

[1] R. W. M. W. J. A. Nelder, “Generalized Linear Models,” *Royal Statistical Society*, p. 370–384, 1972.

[2] P. McCullagh, Generalized linear models, 2nd ed., Routledge, 2019.

[3] D. L. S. a. S. R. Hosmer Jr, Applied logistic regression, John Wiley & Sons, 2013.

[4] D. Collett, Modelling binary data, CRC press, 2002.

[5] M. J. Crawley, The R book, John Wiley & Sons, 2012.

[6] A. a. W. T. Guisan, “Predicting species distribution: offering more than simple habitat models,” *Ecology letters*, pp. 993-1009, 2005.

[7] D. S. e. a. Johnson, “Using Generalized Linear Models to Identify Drivers of Bird Collision Risk at Wind Facilities,” *Journal of Applied Ecology,* pp. 850-860, 2017.

[8] A. J. &. B. Dobson, An introduction to generalized linear models, 2008.

[9] A. Agresti, Categorical Data Analysis, Wiley, 2002.

[10] M. H. N. C. J. N. J. &. L. W. Kutner, Applied Linear Statistical Models, McGraw-Hill, 2004.

[11] A. &. H. J. Gelman, “Data Analysis Using Regression and Multilevel/Hierarchical Models,” *Cambridge University Press,* 2007.

[12] T. T. R. &. F. J. Hastie, The Elements of Statistical Learning: Data Mining, Inference, and Prediction., Springer, 2009.

[13] D. R. Cox, Principles of Statistical Inference, Cambridge University Press, 2006.

[14] A. K. P. H. L. M. J. V. S. &. A. S. D. Malik, “Machine learning for prediction of air quality indices: An overview and future outlook,” *Environmental Pollution,* no. 113034, p. 254, 2019.

[15] P. &. S. V. P. Mishra, “Machine learning techniques for estimation of suspended sediment concentration in rivers.,” *Environmental Modelling & Software,* no. 117, pp. 73-88, 2019.

[16] E. G. &. P. G. P. Hertwich, “Carbon footprint of nations: A global, trade-linked analysis,” *Environmental Science & Technology,* 2009.

[17] I. E. A. (IEA)., CO2 Emissions from Fuel Combustion 2021 Highlights, IEA Publications, 2021.

[18] IPCC, “ Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas,” in *Intergovernmental Panel on Climate Change*, 2019.

[19] A. E. A. S. A. A. &. H. A. Hassan, “Time series analysis and forecasting of CO2 emissions in Saudi Arabia.,” *Sustainability,* no. 4658, p. 10(12), 2018.

[20] S. Z. Y. &. Z. X. Wang, “Short-term carbon dioxide emissions forecasting using a hybrid ARIMA–GARCH–SVM model,” *Energy,* no. 148, pp. 506-521, 2018.

[21] M. H. Q. M. A. T. A. K. &. L. N. C. Shahbaz, “Economic growth, energy consumption, financial development, international trade and CO2 emissions in Indonesia.,” *Renewable and Sustainable Energy Reviews,* no. 25, pp. 109-121, 2013.

[22] Y. M. J. S. Y. &. L. W. Zhang, “CO2 emission forecasting in China based on a hybrid model of least squares support vector machine (LSSVM) and the fruit fly optimization algorithm,” *Sustainability,* no. 10(9), p. 3050, 2018.

[23] J. &. W. C. Xie, “CO2 emissions prediction based on LSTM neural network algorithm.,” *Sustainability,* no. 8625, 2020.

[24] S. S. R. E. &. K. A. Hosseini, “Short-term carbon dioxide (CO2) emissions forecasting using autoregressive integrated moving average (ARIMA) in the transportation sector.,” *Environmental Science and Pollution Research,* no. 31, 2018.

[25] G. E. P. &. J. G. M. Box, “Time Series Analysis: Forecasting and Control,” 1970.

[26] Z. &. C. H. Wang, “A Short-Term Forecasting Model for Carbon Emissions in China Based on ARIMA-GARCH,” *Energy Economics,* 2019.

[27] L. L. P. a. F. L. Ning, “Forecast of China’s carbon emissions based on Arima method,” *Discrete Dynamics in Nature and Society* , pp. 1-12, 2021.

[28] H. J. P. &. K. N. Gupta, “Modeling and forecasting of carbon dioxide (CO2) emissions in Illinois using autoregressive integrated moving average (ARIMA) models,” *Sustainability,* vol. 11, no. 10, p. 2861, 2019.

[29] S. P. M. &. T. A. Kumar, “Forecasting carbon dioxide (CO2) emissions using autoregressive integrated moving average (ARIMA) model,” *Journal of King Saud University-Engineering Sciences,* 2019.

[30] S. W. S. C. &. H. R. J. Makridakis, “Forecasting: Methods and Applications,” *John Wiley & Sons,* 1998.

[31] I. P. o. C. Change, “Sixth Assessment Report,” [Online]. Available: https://www.ipcc.ch/assessment-report/ar6/.

[32] M. R. a. P. R. Hannah Ritchie, “OurWorldInData.org,” OurWorldInData, 2022. [Online]. Available: https://ourworldindata.org/energy.