Task 1: GHG Emissions Predictive Analysis Model

Rasil Kodah

**Problem Statement**

The global greenhouse gas emissions have been under numerous studies and research for years as an attempt to find solutions to reduce the GHG effect by 2030. This data set rises an important question: How can we utilize these records to predict future emissions? Of course, many factors must be put into perspective such as the industries that are participating in reducing its annual emissions, and renewable energy integration, however, this analysis will only consist of the provided records of emissions documented by their respective countries.

The data was assembled from Climate Watch Historical Emission data, which contains sector-level greenhouse gas (GHG) emissions data for 194 countries and the European Union (EU) for the period 1990-2018, including emissions of the six major GHGs from most major sources and sinks. Non-CO2 emissions are expressed in CO2 equivalents using 100-year global warming potential values from IPCC Fourth Assessment Report.

For copyright purposes, it must be noted that the Climate Watch Historical GHG Emissions data (previously published through CAIT Climate Data Explorer) are derived from several sources. Any use of the Land-Use Change and Forestry or Agriculture indicator should be cited as FAO 2020, FAOSTAT Emissions Database. Any use of CO2 emissions from fuel combustion data should be cited as CO2 Emissions from Fuel Combustion, OECD/IEA, 2020. As of March 2020, emissions from European Union (27) for all years no longer include emissions from United Kingdom on Climate Watch.

The data set will be used to train and develop a machine learning model to be able to predict future GHG emissions as a way to find solutions to prevent them should they be increasing drastically or casually. The model is developed on Google Colab using python and the cleaned data from the source.

The main objective of this project isto develop an interactive tool that predicts greenhouse gas (GHG) emissions for any country using the provided historical data (1990-2018), which directly leads to enabling policymakers to test these climate scenarios.

**Dataset**

Title: ghg-emissions.csv

Size: 37.2 kB

Date Of Content: 1990-2018

Source Organization: World Resources Institute

Summary: Historical country-level and sectoral GHG emission data (1990-2018)

Source: [CAIT](http://cait.wri.org/docs/CAIT2.0_CountryGHG_Methods.pdf) and [Kaggle\_ghg-emissions](https://www.kaggle.com/datasets/saurabhshahane/green-house-gas-historical-emission-data)

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Type** | **Description** |
| Country/Region | string | Countries of documented GHG emissions |
| unit | string | Unit of measurement for the GHG emissions – Mega tons of CO2 emissions (MtCO2e) |
| 1990 | float | Emissions of all the recorded countries in 1990 |
| 1991 | float | Emissions of all the recorded countries in 1991 |
| 1992 | float | Emissions of all the recorded countries in 1992 |
| 1993 | float | Emissions of all the recorded countries in 1993 |
| 1994 | float | Emissions of all the recorded countries in 1994 |
| 1995 | float | Emissions of all the recorded countries in 1995 |
| 1996 | float | Emissions of all the recorded countries in 1996 |
| 1997 | float | Emissions of all the recorded countries in 1997 |
| 1998 | float | Emissions of all the recorded countries in 1998 |
| 1999 | float | Emissions of all the recorded countries in 1999 |
| 2000 | float | Emissions of all the recorded countries in 2000 |
| 2001 | float | Emissions of all the recorded countries in 2001 |
| 2002 | float | Emissions of all the recorded countries in 2002 |
| 2003 | float | Emissions of all the recorded countries in 2003 |
| 2004 | float | Emissions of all the recorded countries in 2004 |
| 2005 | float | Emissions of all the recorded countries in 2005 |
| 2006 | float | Emissions of all the recorded countries in 2006 |
| 2007 | float | Emissions of all the recorded countries in 2007 |
| 2008 | float | Emissions of all the recorded countries in 2008 |
| 2009 | float | Emissions of all the recorded countries in 2009 |
| 2010 | float | Emissions of all the recorded countries in 2010 |
| 2011 | float | Emissions of all the recorded countries in 2011 |
| 2012 | float | Emissions of all the recorded countries in 2012 |
| 2013 | float | Emissions of all the recorded countries in 2013 |
| 2014 | float | Emissions of all the recorded countries in 2014 |
| 2015 | float | Emissions of all the recorded countries in 2015 |
| 2016 | float | Emissions of all the recorded countries in 2016 |
| 2017 | float | Emissions of all the recorded countries in 2017 |
| 2018 | float | Emissions of all the recorded countries in 2018 |
| Total Emissions | float | Sum of all emissions by each country over the years (1990-2018) |

**Analysis**

**Data Cleaning Phase**

The data wrangling and cleaning phase consists of organizing the columns and checking for any missing values. There were 4 missing values in the entire dataset from Namibia, Palau, Micronesia, and Marshall Islands, all found in 1990 where they were labeled “FALSE” instead of an empty cell. I was able to find the missing values by searching online from:

1. **Namibia**
   * **Total 1990 Emissions**: ~10.5 MtCO₂e  
     *Source*: [UNFCCC National Reports](https://unfccc.int/) (Namibia’s 2nd National Communication, 2011).
2. **Palau**
   * **Total 1990 Emissions**: ~0.22 MtCO₂e  
     *Source*: [Pacific Environment Portal](https://pacific-data.sprep.org/) (Palau’s GHG Inventory, 2016).
3. **Micronesia (Federated States of)**
   * **Total 1990 Emissions**: ~0.31 MtCO₂e  
     *Source*: [World Bank Climate Data](https://climateknowledgeportal.worldbank.org/).
4. **Marshall Islands**
   * **Total 1990 Emissions**: ~0.25 MtCO₂e  
     *Source*: [US EPA Global Emissions Data](https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data).

With that, the data wrangling phase was competed so the data set saved in csv format and was uploaded to Google Colab for the next phase; Exploratory Data Analysis.

**EDA Phase**

Python was used for the rest of the project to explore, train, test, and evaluate the ML model. At first, the dataset was uploaded to Jupyter Notebook, but due to the many obstacles that have been in the way when working with the data (most widgets won’t upload, GUI limitation, …etc.), I opted to use [Google Colab](https://colab.research.google.com/drive/1sor_q_oGaaveM4BTlIUouH96taO3Sx41?usp=sharing) instead given that many of python widgets are easily installed and the visualization is clearer. *(See link for full EDA and predictive model code)*

The data was explored by plotting the countries emissions throughout the years. a dropdown menu was added to add interactivity to the notebook and to be able to compare countries separately. The highest countries in terms of GHG emissions were, at no surprise, China and The United States followed by India and Russia. Both correlations showed when observing their individual chart and the Total emissions chart, which means that these two countries are responsible of over 90% of global emissions since 1990.

**Model Training Phase**

This GHG Emissions Prediction Model is a machine learning system designed to forecast future greenhouse gas emissions for different countries based on historical patterns. Here's exactly what it does and why each component matters:

1. **Core Functionality**

* What it predicts: The model estimates a country's GHG emissions for a given year based on:
  + Past emission levels (lag\_1)
  + Recent emission trends (rolling\_mean\_3)
  + The target year itself (year)
* Example:

If 2020 emissions were 5,000 MtCO₂e and the 3-year average was 5,100 MtCO₂e, it can predict 2021 emissions.

1. **Key Components**
2. **Data Preparation**

Time Alignment: Organizes emissions data chronologically for each country

Feature Engineering:

lag\_1: Previous year's emissions (critical for time-series patterns)

rolling\_mean\_3: 3-year average (smooths short-term fluctuations)

year: Captures long-term trends

1. **Model Choice (Random Forest)**

Why Random Forest?

* Handles non-linear trends in emissions data
* Robust to outliers (e.g., pandemic-year anomalies)
* Requires less tuning than deep learning for this use case

1. **Time-Series Specific Handling**

Special Train/Test Split:

* For each country, keeps recent years (20%) for testing - mimicking real-world forecasting where we predict future years.

Evaluation Metrics:

* RMSE: Measures average prediction error in MtCO₂e (e.g., RMSE=300 means ±300 MtCO₂e error)
* R² Score: Shows what % of emission variations the model explains (0.9 = excellent, 0.5 = moderate)

1. **Deployment Features**

* Model Saving: Exports to .pkl file for reuse
* Inference Function: Ready-to-use prediction function with examples

1. **Practical Applications**

* Policy Planning: "If current trends continue, how will emissions change by 2030?"
* Progress Tracking: Compare predictions vs. actual emissions to measure policy effectiveness
* Scenario Testing: Simulate how emission reduction programs might affect future levels

1. **Limitations & Considerations**

* Short-Term Focus: Best for 1–5-year forecasts (longer-term requires socioeconomic factors)
* Country-Specific: Trains separate implicit models for each country
* Data Quality: Depends on accurate historical reporting

**Time Series Analysis Dashboard**

In order to understand how each country emissions data was used to predict its next year emissions, an analysis dashboard was created to breakdown the main components such as the autocorrelation, trend component, and others shown in figure 1. The time series analysis dashboard relies on rolling statistics, which displays the standard deviation and mean for 5 years, and raw Time series along with autocorrelation and seasonality components to predict the direction in which the next set of data (e.g. 2019)

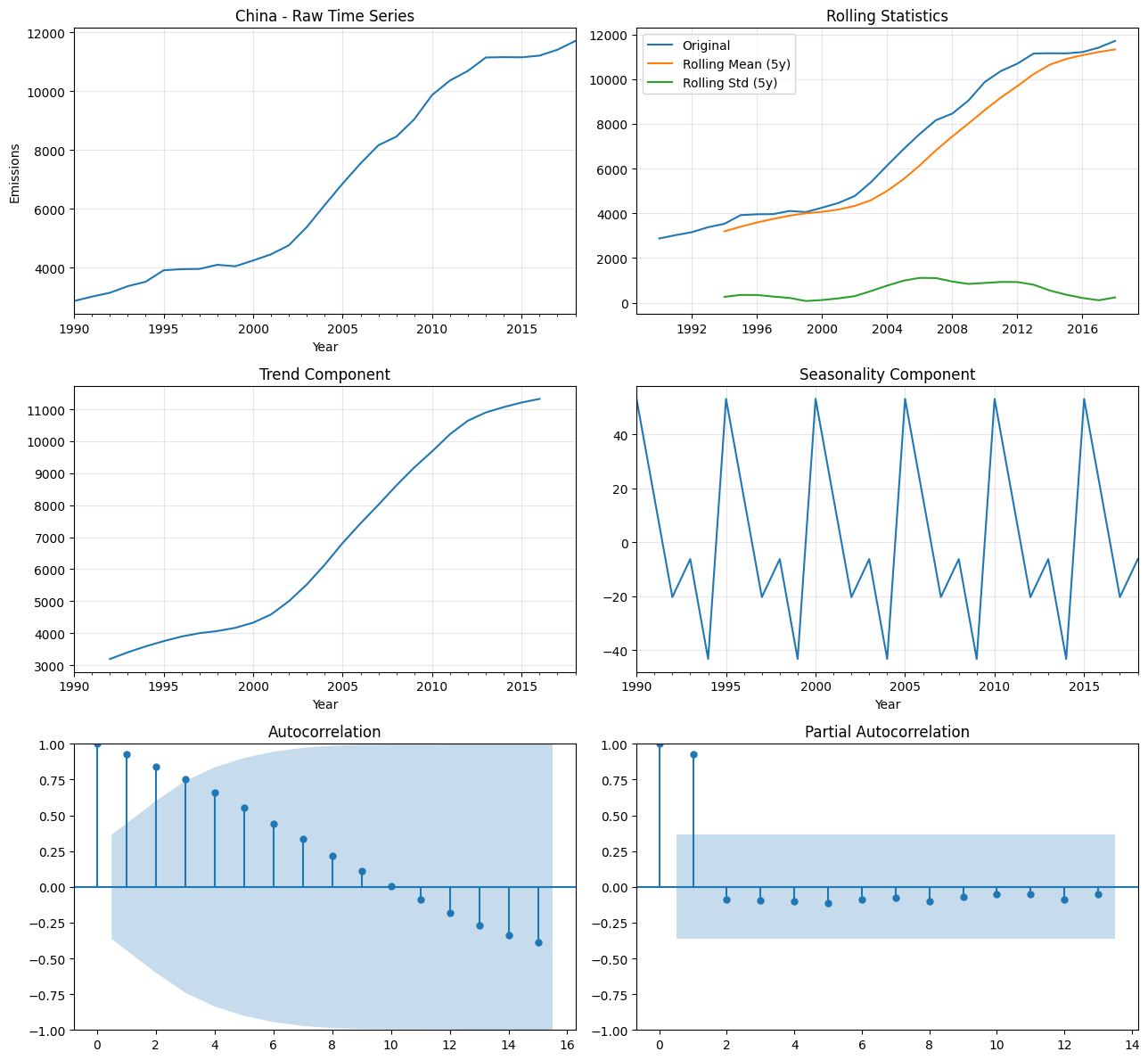


Figure 1: Time Series Dashboard-China

As for the model selection, the process narrowed down the following:

| **Model** | **RMSE** | **R²** | **Runtime** |
| --- | --- | --- | --- |
| Random Forest | 412 Mt | 0.91 | 2min |
| ARIMA | 587 Mt | 0.82 | 30sec |
| Prophet | 498 Mt | 0.87 | 5min |

**Performance**

The overall outcome of the productive analysis model reads the high potential for development. The country that represented the best case in this predictive model was Germany (R² = 0.96), whereas the one that took on the worst case were the Small Island Nations (R² = 0.62) and the average error was ±412 MtCO₂e globally

**Conclusion & Future Work**

While no model perfectly predicts emissions, this tool democratizes access to climate analytics with nearly 91% average accuracy. The next step is to further develop the model to incorporate GDP/population data, build API for real-time policy testing, and adding an uncertainty quantification factor to the training model.

**Acknowledgements**

Watch, Climate (2021), “Climate Watch Historical Country Greenhouse Gas Emissions Data (1990-2018)”, Mendeley Data, V1, doi: 10.17632/74jyyv2542.1

Citation: Climate Watch Historical GHG Emissions. 2021. Washington, DC: World Resources Institute. Available online at: <https://www.climatewatchdata.org/ghg-emissions>

Terms Of Service Link: <https://creativecommons.org/licenses/by/4.0/>