DATA 1030 Final Report
Predicting Rwanda CO2 Emission

Long Hei Chan
Brown University
Github Link: https://github.com/clhperry/rwanda_co2

1. Introduction

1.1. Motivation

Accurately tracking carbon emissions stands as a crucial measure in the battle against climate change. It provides invaluable insights for researchers and governments into the origins and trends of carbon output. While Europe and North America have robust systems for monitoring emissions on land, there's a significant scarcity of such systems across Africa.

The challenge at hand involves developing machine learning models that leverage open-source data from Sentinel-5P satellite observations to forecast forthcoming carbon emissions. [1] These solutions could prove instrumental in helping governments and various stakeholders estimate emission levels throughout Africa, particularly in areas where direct on-site monitoring remains unfeasible. [2]

The dataset presents a time-series regression problem for predicting future CO2 emissions in Rwanda. Emission data is provided from 2019 to 2021.

1.2. Previous Work

As this is a dataset from Kaggle, there is a leaderboard measured by RMSE score. Most models are trained with data from 2019 to 2021 and tested with data from 2022, with top scores performing at a score of around 17 to 18.

However, due to licensing issues, Kaggle stopped providing data for 2022 in Nov 2023, thus new submissions can only train from 2019 to 2020 and test with 2021.

2. Explanatory Data Analysis

2.1. Target Variable Analysis

To understand the target variable, a histogram plot is created to understand its distribution across different locations. As shown below, the target variable is right-skewed and follows a long tail distribution (Figure 1).

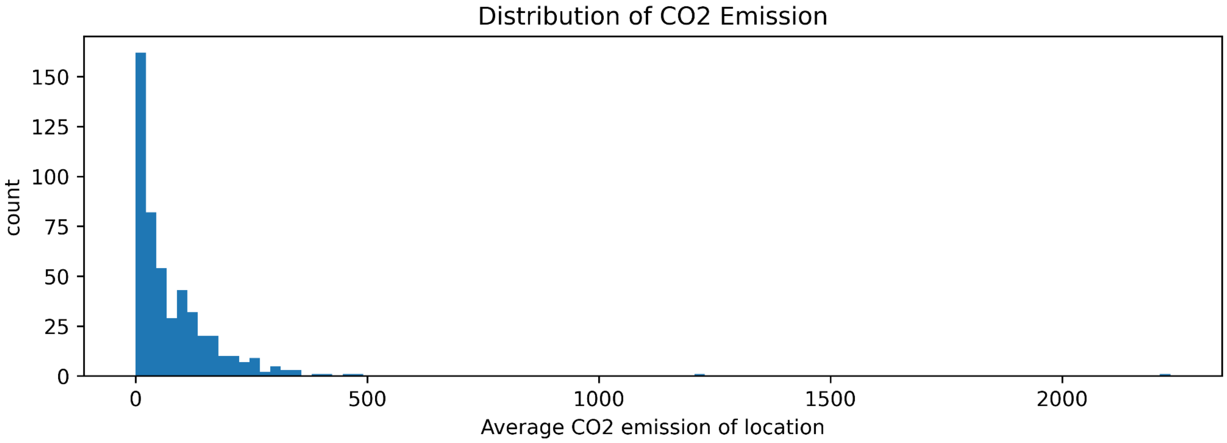

Figure 1 - Distribution of average CO2 emission (target variable) against location

2.2 Seasonality

As this is a time series data, analysis is also done between CO2 emissions of different locations against time (Figure 2). It can be observed that there is an annual seasonality and there are two locations with exceptionally high emissions.

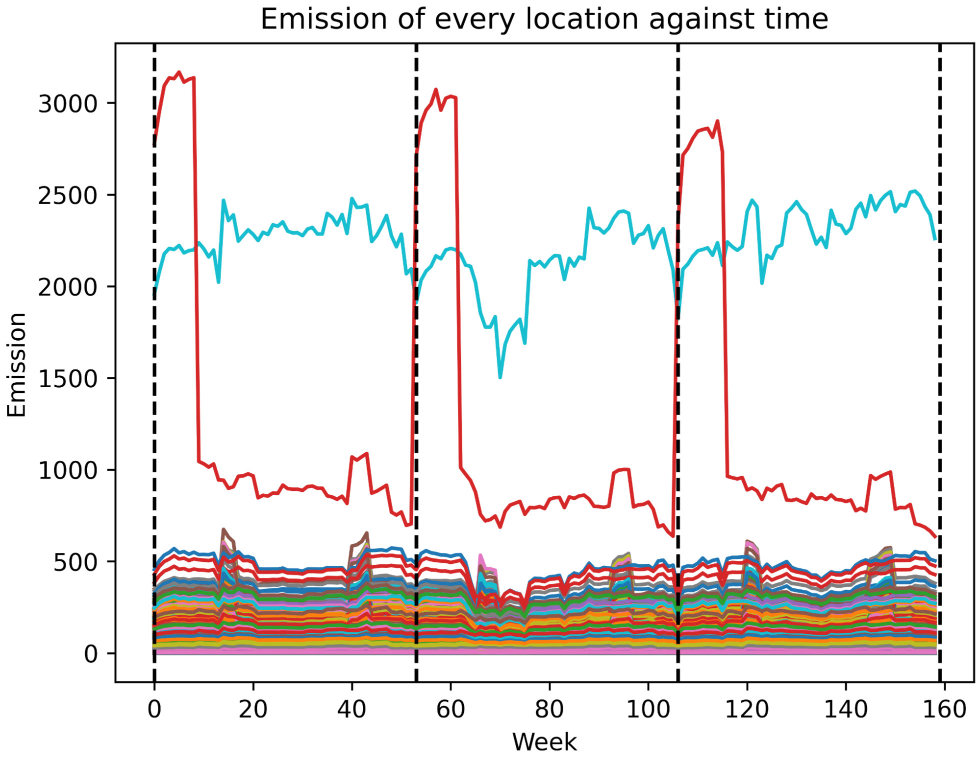

Figure 2 - CO2 emission of every location against time

2.3. Missing Values

In this dataset, an interesting observation is that most data related to UV is missing (Figure 3), thus these features may need to be dropped entirely.

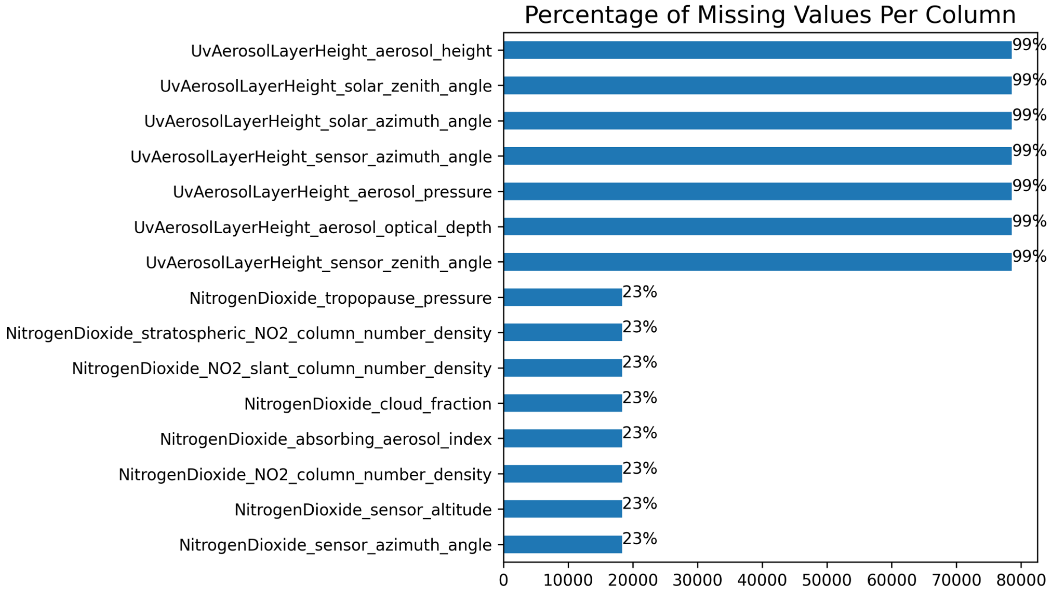

Figure 3 - Percentage of Missing Values (Top 15)

2.4. Correlation of gases

As the dataset contains the density of multiple gases other than CO2 (e.g. carbon monoxide, ozone), a correlation heatmap is plotted to understand their relationship (Figure 4). CO2 emission and the density of other gases are very weakly correlated, thus these data are highly likely to be noise and would not contribute to the model.

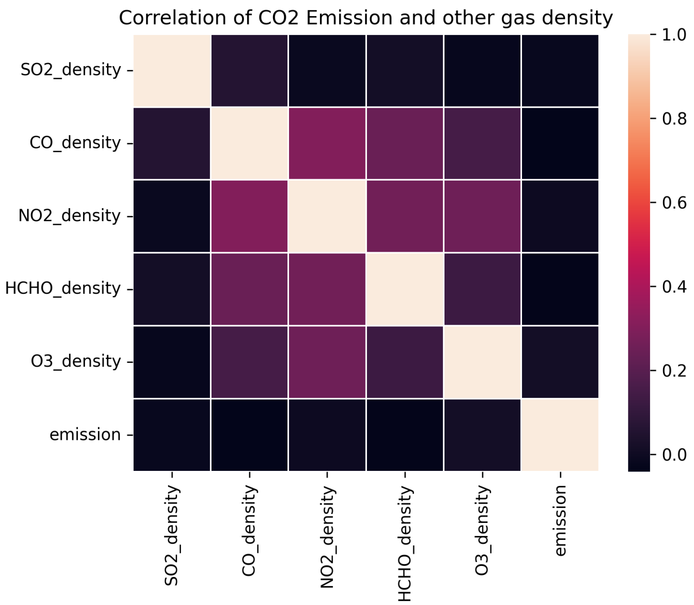

Figure 4 - Correlation Heatmap between CO2 emission and density of other gases

3. Methods

3.1. Splitting Strategy

As this is a time series data with annual seasonality, I used the earlier years as the training set and the last year as the test set, aligning with this project's goal to predict the future with existing data. For cross-validation, I utilized LeaveOneGroupOut by grouping the data according to their year.

3.2. Data Preprocessing

Due to the noisy nature of gas-related data, they are dropped to improve model efficiency and reduce variance.

As CO2 emissions of locations are likely to be affected by nearby locations, I used clustering methods to engineer two additional features, which are the group of the location and the distance to the highest emission location.

3.3. Pipeline

Four machine learning algorithms are used in this analysis: Random Forest Regression, XGBoost Regression, Elastic Net Regression, and KNN Regression. For each algorithm, several hyperparameters were tuned. (Table 1)

Model	**Tuned Parameters**	**Best CV RMSE (Mean ± std)**
Elastic Net Regression	alpha: [**0.1**, 1.0, 10.0], l1_ratio: [0.1, 0.5, 0.7, **0.9**]	132.46 ± 5.01
KNeighborsRegressor	n_neighbors: [**3**, 5, 7, 9], p (distance metric): [**1**, 2]	29.60 ± 0.11
XGBoost	n_estimators: [100, **200**, 300], max_depth: [7, **9**, 11], learning_rate: [**0.1**, 0.01]	28.93 ± 0.26
Random Forest	n_estimators: [100, **200**, 300], max_depth: [**None**, 5, 10]	29.13 ± 0.13

Table 1 - Tuned parameter values for each Machine Learning algorithm

3.4. Metrics

The metric used to evaluate the models' performance is RMSE, enabling the performance of the model to be more interpretable, as the error value would be in the same unit as the target variable.

In the cross-validation, tree-based modeling and KNN algorithm have similar performance, while the Elastic Net Regression performed significantly worse.

4. Results

4.1. Baseline Score

A simple moving average with a window size of 3 is used as the baseline model for the time series data, and the score is 26.35.

4.2. Model Performance

From the comparison of different models' test scores, we can see that Random Forest is the most predictive as it has the lowest RMSE score, while KNN outperforms the baseline the most with regard to the standard deviation.

Model	Mean RMSE test score	Standard deviation	(Baseline - RMSE test score) / standard deviation
Random Forest	19.11	0.13	55.69
XGBoost	19.88	0.26	24.88
Elastic Net	135.23	5.01	-21.73
KNN	19.28	0.11	64.27

Table 2 - Mean values of each Machine Learning algorithm

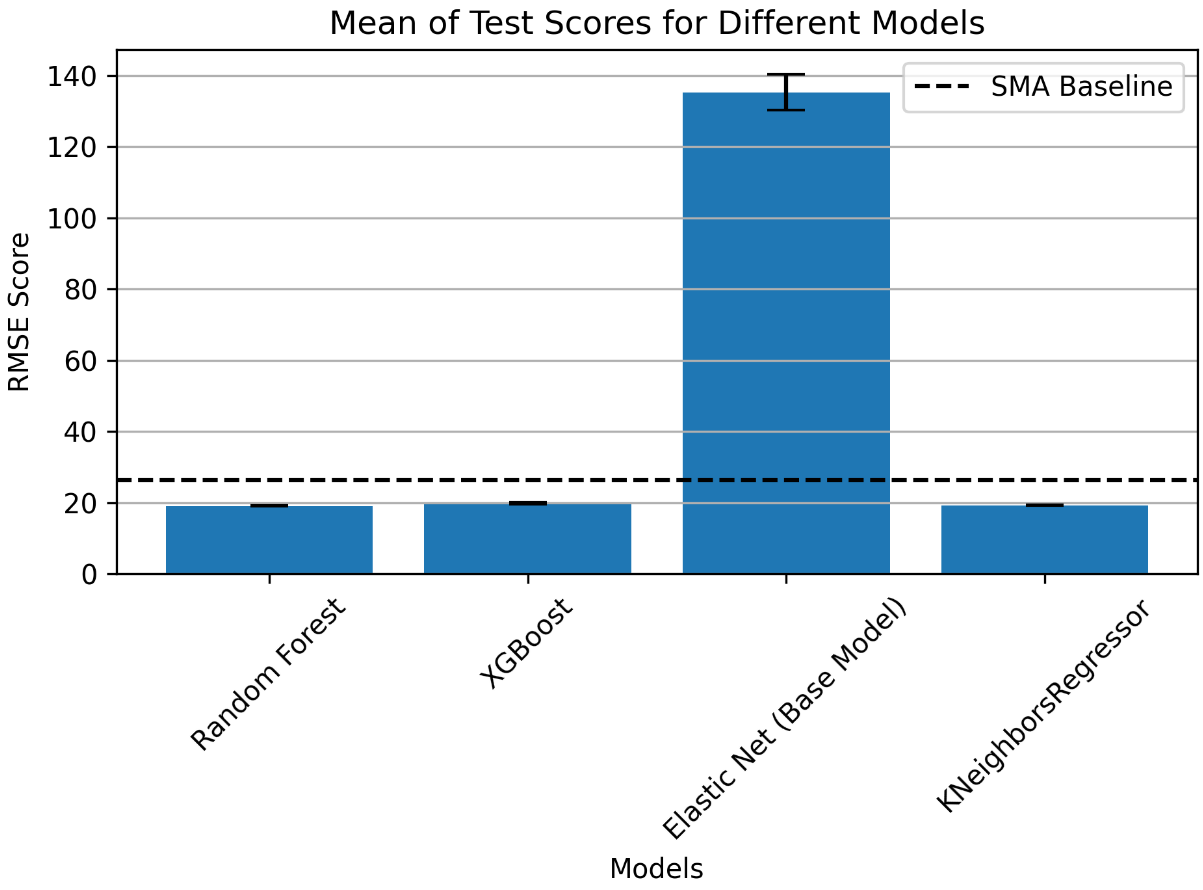

Figure 5 - Model performance and comparison to baseline score

The Random Forest model is also able to capture the seasonality across the test set (Figure 6), as it accurately predicted the two spikes during the year of 2021while following the overall trend nicely.

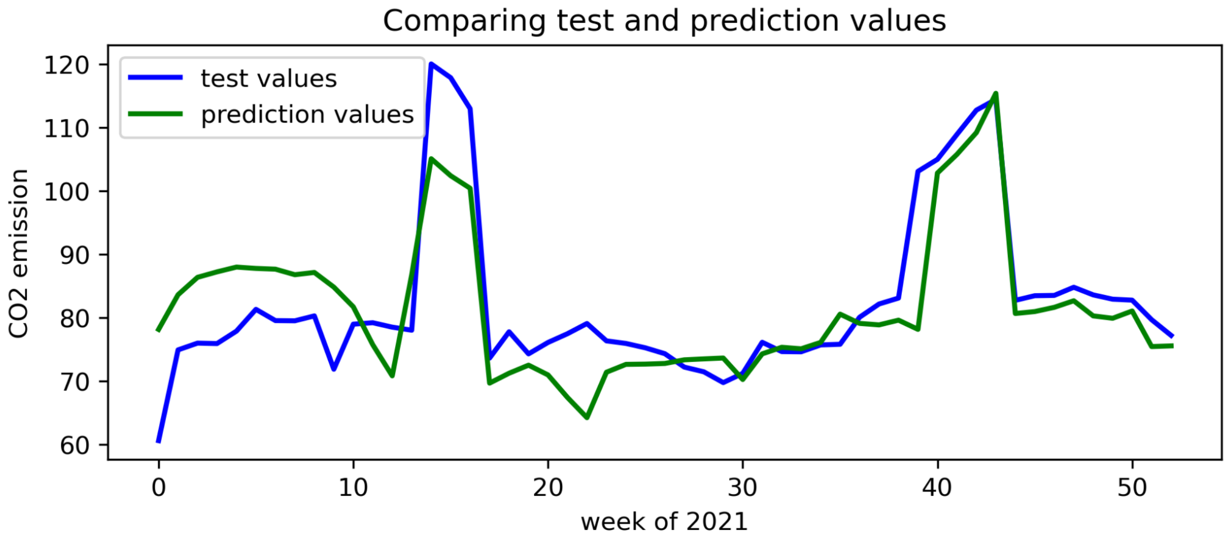

Figure 6 - Comparison of test set values and predicted values

4.3. Global Feature Importance

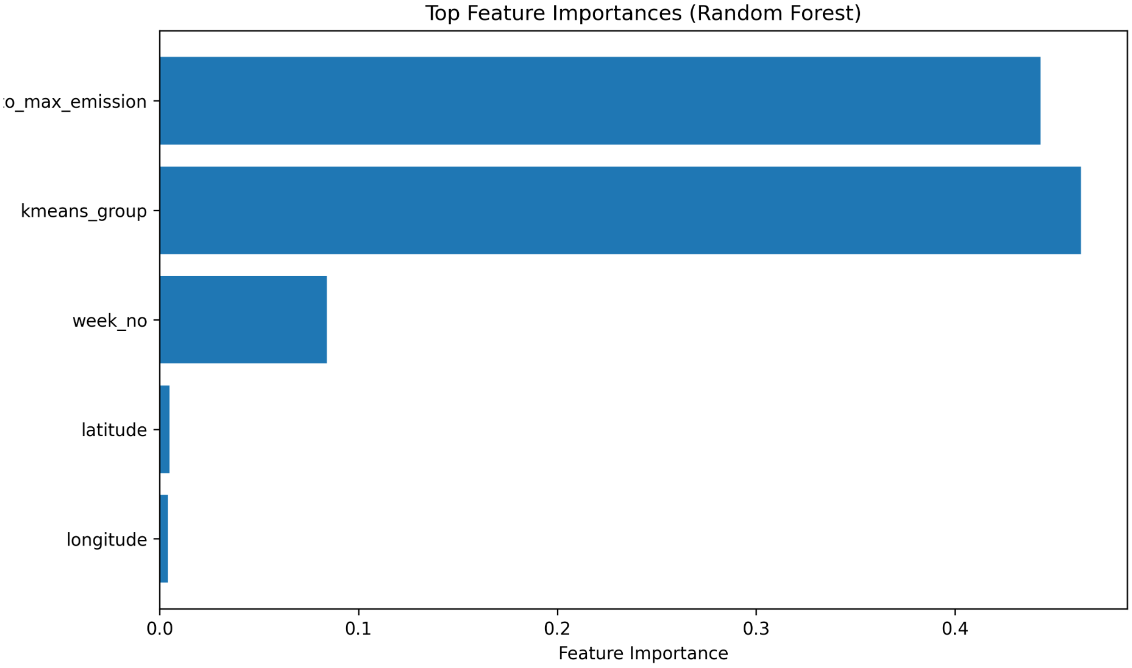

Figure 7 - Top Feature Importances of Random Forest model

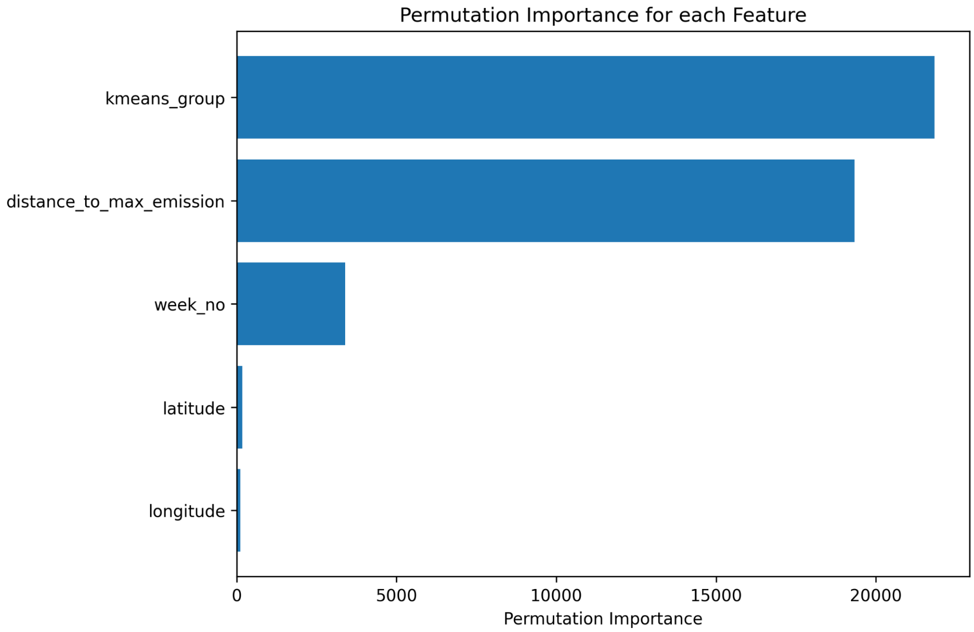

Figure 8 - Top Permutation Importances of Random Forest model

From the above global feature importance graphs, we can see that the most important features are the engineered features of the cluster group that the location belongs to and the distance to the location with the maximum emission, followed by the time of the year.

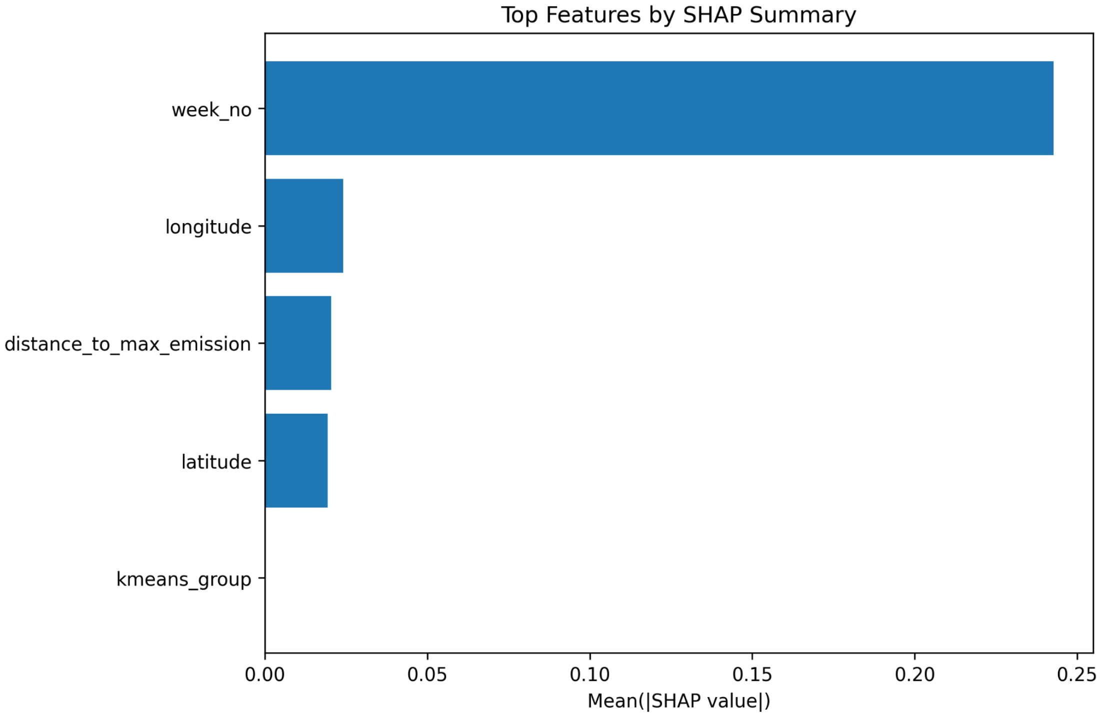

Figure 9 - SHAP values of different features

As for the SHAP values summary, the time of the year has a more significant impact, but there is a caveat where only the first 100 rows are sampled due to computational limits, as more rows would cause the kernel to reset.

4.4. Local Feature Importance

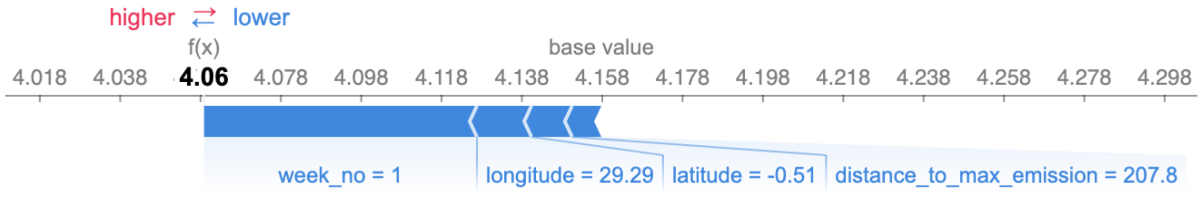

Figure 10 - SHAP value for local index 1

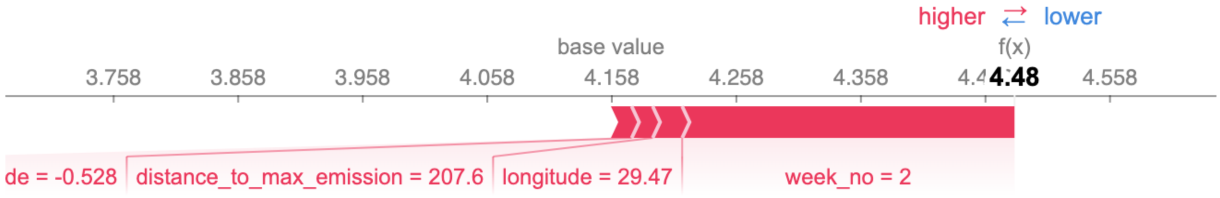

Figure 11 - SHAP value for local index 55

For the above local indices, they share similar characteristics where the number of the week contributes the most to the emission value, followed by other features such as the location.

5. Outlook

5.1. Possible Improvements

Due to the COVID-19 outbreak, Rwanda went into lockdown in 2020. As the training data includes the CO2 emissions in 2020, there could be possible improvements by adjusting the emission data to match normal years.

Also, as Rwanda is a developing country, there is likely an overall upward trend of CO2 emissions due to increased economic activities. Due to the short timeframe of this dataset and the COVID-impacted year, this trend may not be reflected accurately.

5.2. Additional Data

As the dataset only contains three years of data, including a year of COVID-19, it is difficult for the model to generalize well enough for future data. Therefore, more data from the past would be really helpful for the model to perform better.

To further improve the model, obtaining data related to other factors that could affect CO2 emission would be helpful, including predicted economic growth of different regions, population growth, and energy demand. [3]

6. References

1. *Predict CO2 emissions in Rwanda | Kaggle*. (n.d.). https://www.kaggle.com/competitions/playground-series-s3e20
2. *The battle for Earth's climate will be fought in Africa*. (n.d.). Wilson Center. https://www.wilsoncenter.org/article/battle-earths-climate-will-be-fought-africa
3. Li, S., Siu, Y. W., & Zhao, G. (2021). Driving Factors of CO2 emissions: Further study based on Machine learning. *Frontiers in Environmental Science*, *9*. https://doi.org/10.3389/fenvs.2021.721517

7. Github Repository

https://github.com/clhperry/rwanda_co2