DATA ANALYSIS REPORT:

PREDICTING CO_2 EMISSIONS IN RWANDA

Prepared By:
Kushal Chinthaparthi

Contents

PREDICTING CO_2 EMISSIONS IN RWANDA

Executive Summary

Introduction

The "Predicting CO_2 Emissions in Rwanda" report is a data-driven initiative aimed at addressing the critical issue of carbon dioxide (CO_2) emissions in Rwanda. As global concerns about climate change intensify, this project takes a proactive approach to provide a predictive model for estimating CO_2 emissions in Rwanda, thereby supporting the country's commitment to sustainable development. This executive summary offers a brief overview of the project, identifies the primary audience, and justifies the report's importance to key stakeholders.

Audience

The primary audience for this report includes policymakers, environmental experts, businesses, and organizations invested in Rwanda's and the world in sustainable development. Policymakers rely on accurate data for evidence-based decision-making, environmental experts seek insights for informed interventions, and businesses aim to align their operations with sustainable practices.

Importance to Policymakers:

Policymakers play a pivotal role in shaping legislation, regulations and policies that guide a nation's trajectory. For Rwanda, grappling with the challenges of climate change necessitates informed decision-making. This report provides policymakers with a robust predictive model, derived from historical emission data, socioeconomic indices, and environmental factors. By having a tool that forecasts CO_2 emissions accurately, policymakers can develop and implement targeted policies, fostering sustainable growth while mitigating the impact of climate change.

Importance to Environmental Experts:

Environmental experts and international organizations are crucial in driving initiatives that protect and preserve the environment. This report's significance lies in its capacity to unravel the factors contributing to CO_2 emissions in the world based on the Rwanda's case. With actionable insights derived from rigorous data analysis, environmental experts can tailor interventions to address specific challenges, leading to more effective and targeted environmental conservation efforts.

Importance to Businesses:

Businesses globally are increasingly recognizing the importance of sustainability in their operations. This report serves as a valuable resource for businesses by analyzing the factors contributing to CO_2 emissions. Armed with this information, businesses can make informed decisions to reduce their carbon footprint, implement sustainable practices, and contribute to the global transition to an environmentally conscious and sustainable future.

By seeking to provide accurate predictions and actionable insights, this report empowers policymakers, environmental experts, and businesses to collaboratively address the urgent issue of CO_2 emissions, contributing to sustainable development and global climate change mitigation efforts.

Detailed Report

Problem Description

Considering global climate change, Rwanda, like many other nations, must manage and reduce its carbon dioxide (CO_2) emissions. For informed policy decisions, sustainable development, and the mitigation of the negative effects of climate change, accurate evaluation, forecast, and knowledge of CO_2 emissions are crucial. However, Rwanda currently lacks a thorough predictive model to precisely predict its upcoming CO_2 emissions.

Analytical Statement:

The "Predicting CO_2 Emissions in Rwanda" initiative seeks to solve the issue of insufficient CO_2 emission prediction modelling in Rwanda. This project aims to build a strong machine learning model that can forecast CO_2 emissions in the nation by utilizing historical emission data, socioeconomic indices, and environmental factors. We aim to provide actionable insights to key stakeholders, including policymakers, environmental experts, and businesses, through rigorous data analysis, model selection, and interpretation.

Business Goal:

The primary objective of this data analytics capstone project is to contribute to the global effort in combating climate change by developing a machine learning model to predict future carbon emissions in Africa and the world. By leveraging open-source CO_2 emissions data from Sentinel-5P satellite observations, the project aims to provide accurate insights into carbon mass output across the continent.

This initiative addresses the critical need for monitoring carbon emissions in Africa, where existing on-the-ground monitoring systems are limited. The ultimate business goal is to empower governments and other stakeholders with the tools to estimate carbon emission levels, facilitating informed decision-making and targeted interventions to mitigate the impact of climate change.

Data Analysis Goal:

The primary objective of this is to develop a machine learning model using open-source CO_2 emissions data from Sentinel-5P satellite observations to predict and analyze future carbon emissions in Africa. This involves data exploration, feature engineering, and model optimization to provide accurate insights for governments and stakeholders, enabling effective monitoring and mitigation strategies against climate change and promote sustainable practices.

Data Description

The dataset is encompassing approximately 497 unique locations across diverse areas in Rwanda such as farmlands, urban regions, and power plants, spans from 2019 to 2021 and serves as the training set for building our models. It is available on Kaggle (https://www.kaggle.com/competitions/playground-series-s3e20/overview). Our task extends to predicting CO_2 emissions for the year 2022.

The dataset has seven key features extracted weekly from January 2019 to November 2022, including Sulphur Dioxide, Carbon Monoxide, Nitrogen Dioxide, Formaldehyde, UV Aerosol Index, Ozone, and Cloud. Each of these features has associated sub-features like “**column_number_density,”** representing the vertical column density at ground level, calculated using the DOAS technique. These features will play a pivotal role in training our machine learning models to predict CO_2 emissions.

Below is a detailed breakdown of the features:

1. **Sulphur Dioxide (SO2)**: [COPERNICUS/S5P/NRTI/L3_SO2]: Measures the concentration of Sulphur Dioxide in the atmosphere.
2. **Carbon Monoxide (CO)**: [COPERNICUS/S5P/NRTI/L3_CO]: Measures the concentration of Carbon Monoxide in the atmosphere.
3. **Nitrogen Dioxide (NO2)**: [COPERNICUS/S5P/NRTI/L3_NO2]: Measures the concentration of Nitrogen Dioxide in the atmosphere.
4. **Formaldehyde (HCHO)**: [COPERNICUS/S5P/NRTI/L3_HCHO]: Measures the concentration of Formaldehyde in the atmosphere.
5. **UV Aerosol Index**: [COPERNICUS/S5P/NRTI/L3_AER_AI]: Indicates the presence of aerosols in the atmosphere using ultraviolet light measurements.
6. **Ozone (O3)**: [COPERNICUS/S5P/NRTI/L3_O3]: Measures the concentration of Ozone in the atmosphere.
7. **Cloud**: [COPERNICUS/S5P/OFFL/L3_CLOUD]: Provides cloud-related parameters derived from satellite observations.

The sub-feature “**column_number_density”** is common across the features, representing the vertical column density of the respective gases or particles at ground level.

The dataset, rich in atmospheric observations, provides a solid foundation for developing machine learning models aimed at predicting CO_2 emissions, contributing significantly towards environmental monitoring and policymaking in Rwanda.

Data Preparation

Implementation Steps:

Data Collection and Preparation:

Collecting a comprehensive historical perspective on Rwanda's CO_2 emissions involved the compilation of data from government reports and environmental organizations. Supplementary information on population dynamics, energy consumption, industrial activities, transportation patterns, and land use was also acquired to provide a holistic context. Rigorous data quality tests were conducted, addressing issues such as missing values and outliers, while ensuring a structured and standardized format, typically in CSV or Excel, for enhanced clarity and consistency in the dataset.

Exploratory Data Analysis (EDA):

The data was visually analyzed using graphs, histograms, scatter plots, and correlation matrices, with trends and relationships between factors such as population, energy use, and industrial activity with CO_2 emissions in Rwanda being revealed.

Feature Selection and Engineering:

1. **Statistical Tests:** These are methods used to analyze data and draw conclusions. They help us understand relationships, patterns, and significance. For feature selection, statistical tests can identify which features (variables) are most relevant for predicting an outcome.
2. **Domain Expertise:** Refers to knowledge and understanding of a specific field or subject. Domain experts contribute insights about which features matter most based on their expertise. Their input helps guide feature selection and model building.
3. **Feature Engineering:** Involves creating new features from existing ones or transforming existing features. The goal is to improve model performance by providing more relevant information. Examples include creating ratios, aggregating data, or encoding categorical variables.

4. **Enhancing Model Performance:** By combining statistical tests, domain expertise, and feature engineering, we create a better representation of the data. This leads to improved model accuracy, generalization, and predictive power.

Model Selection and Development:

In our analysis we have selected three machine learning models which we believe can estimate our future predictions and the output of these models justify the future predictions which will support necessary action to control carbon emissions.

LightGBM is a powerful gradient boosting framework that is widely used for machine learning tasks. It is particularly popular for its efficiency, scalability, and ability to handle large datasets. Here are some key points about LightGBM:

1. **Gradient Boosting Algorithm:** LightGBM is based on the gradient boosting algorithm, which combines weak learners (usually decision trees) to create a strong predictive model. It builds trees sequentially, with each tree correcting the errors of the previous ones.
2. **Leaf-wise Growth:** Unlike traditional depth-wise growth, LightGBM grows tree leaf-wise. It chooses the leaf with the maximum reduction in loss during each split.
3. **Histogram-Based Approach:** LightGBM uses histograms to bin feature values, reducing memory usage and speeding up training.
4. **Categorical Feature Handling:** It efficiently handles categorical features without one-hot encoding.
5. **Gradient-Based One-Side Sampling (GOSS):** GOSS focuses on the samples with large gradients, improving training speed.
6. **Exclusive Feature Bundling:** It bundles exclusive features together for better representation.
7. **Regularization:** LightGBM supports L1 (Lasso) and L2 (Ridge) regularization.
8. **Early Stopping:** It stops training when the validation loss stops improving.
9. **Hyperparameters:** LightGBM has various hyperparameters to control model behavior, such as learning rate, maximum depth, number of leaves, and minimum data in leaves. Hyperparameter tuning is crucial for optimal performance.

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm that has gained popularity for its performance in various tasks. We explored some key aspects of XGBoost:

1. **Gradient Boosting Algorithm:** XGBoost is an ensemble learning method based on gradient boosting. It sequentially builds a strong model by combining multiple weak models (usually decision trees).
2. **Features of XGBoost:** Regularization: XGBoost includes L1 (Lasso) and L2 (Ridge) regularization terms to prevent overfitting.
3. **Custom Loss Functions:** You can define custom loss functions based on your problem.
4. **Handling Missing Values:** XGBoost can handle missing data during training.
5. **Parallelization:** It efficiently parallelizes tree construction.
6. **Tree Pruning:** It prunes trees during training to improve generalization.
7. **Cross-Validation:** XGBoost supports k-fold cross-validation.

Random Forest Machine Learning Model is an ensemble learning technique that combines multiple decision trees to create a robust and accurate model. Here are the key features of Random Forest:

1. **Ensemble of Decision Trees:** Random Forest builds an ensemble of decision trees during training. Each tree is trained on a random subset of the data (bootstrap samples) and a random

subset of features. The final prediction is an average (for regression) or majority vote (for classification) of individual tree predictions.

2. **Bagging (Bootstrap Aggregating):** Random Forest uses bagging to reduce overfitting. By training each tree on a different subset of data, it reduces variance and improves generalization.
3. **Random Feature Selection:** At each split, Random Forest randomly selects a subset of features to consider. This randomness decorrelates the trees and prevents them from becoming too similar.
4. **Out-of-Bag (OOB) Error:** Random Forest estimates its performance using the OOB samples (data not used during training). OOB error provides an unbiased estimate of model accuracy.
5. **Hyperparameters:** Key hyperparameters include "n_estimators", "max_depth", "min_samples_split", & "max_features". Hyperparameter tuning is essential for optimal performance.
6. **Applications:** Random Forest is versatile and works well for both regression and classification tasks. It handles missing values and noisy data effectively.

Hyperparameter Tuning:

Optimized model hyperparameters using techniques like grid search and random search.

Model Evaluation:

Evaluation metrics, including MAE (Mean Absolute Error), MSE (Mean Squared Error), and R2, were employed to assess model performance on the test dataset, and checks for overfitting or underfitting were conducted with necessary actions taken.

Feedback and Iteration:

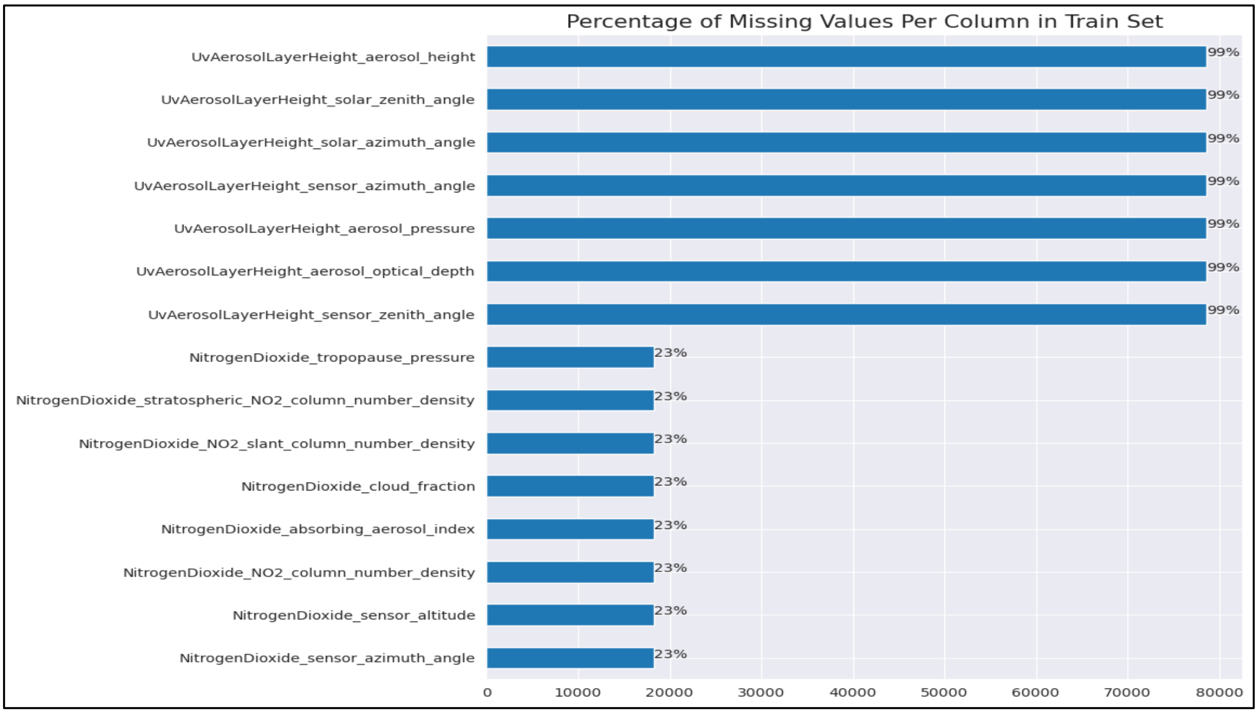
Input from stakeholders and subject-matter experts was gathered to enhance models and recommendations, and regular model updates were incorporated with the inclusion of fresh data.

Knowledge Sharing:

Disseminated project findings, insights, and suggestions to the wider community, including policymakers, researchers, and environmental organizations.

- Unique IDs were generated for each location using latitude and longitude.
- Within the dataset, there were missing data columns. More than 50% of the missing values were dropped with the remaining missing values filled with the meaning of the data columns.
- Only a subset of features was retained with the "Ozone_solar_azimuth_angle" feature being filled forward and backward within the groups.
- New features were created, including rotated coordinates and distances to specific points of interest using the Haversine formula.
- Temporal features such as the month extracted from the year and week number, indicators for COVID-19 related periods, and sinusoidal transformations of the week number were also added.
- Data was preprocessed for both training and test datasets, which were then combined.
- KMeans clustering was applied to the coordinates to create spatial features, and the dataset was then split back into the training and test sets.
- The preprocessing phase included a standardization step where numerical features were normalized using the Standards Caler method, ensuring zero mean and unit variance. This excluded "week_no," "covid_flag," "latitude," "longitude," and "emission" to maintain their original scale for interpretability and context relevance.

- Normalization was uniformly applied across the training and test datasets to maintain consistency in data distribution, a critical factor for the reliable performance of the employed machine learning algorithms.



Data Analysis Solution

To arrive at the Data Analysis solution, we deployed:

1. Time series analysis to identify trends and seasonal patterns.
2. Anomaly detection to pinpoint outliers, such as the significant drop in emissions during the Q2 of 2020 due to the Covid-19 virus outbreak.

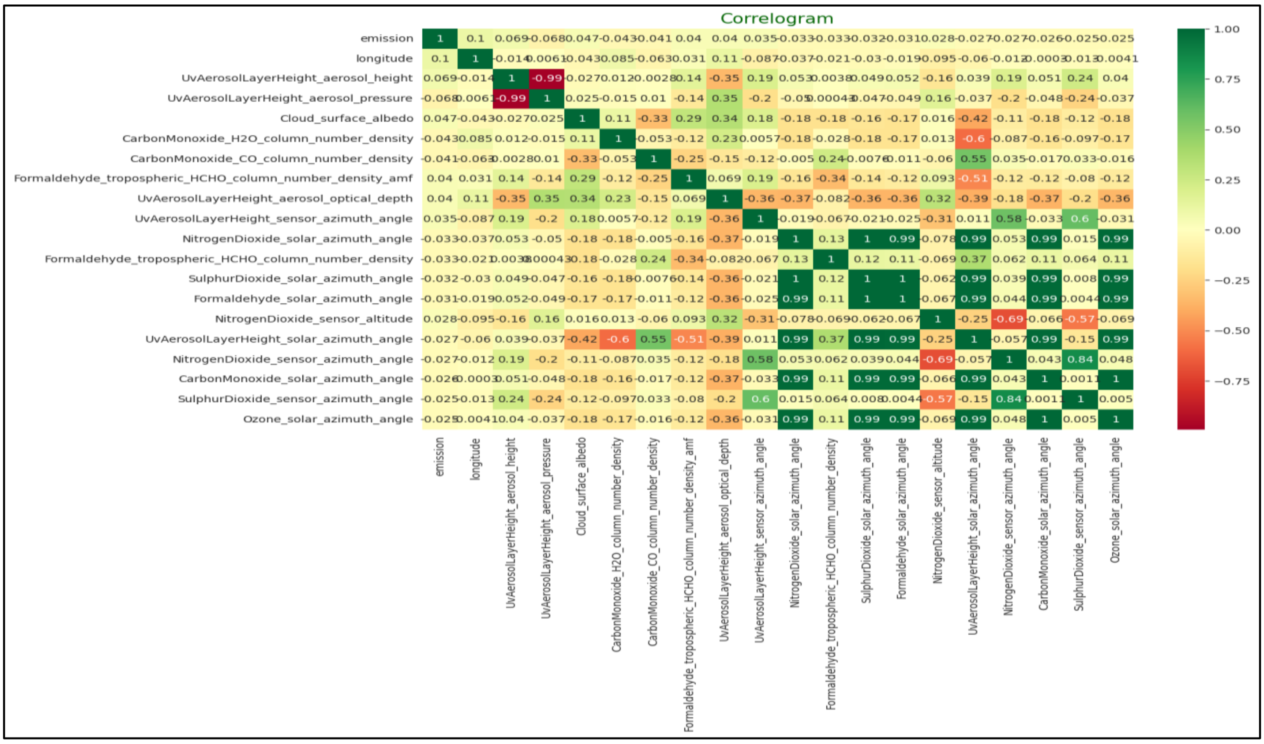


Figure 2.0: Correlation Matrix

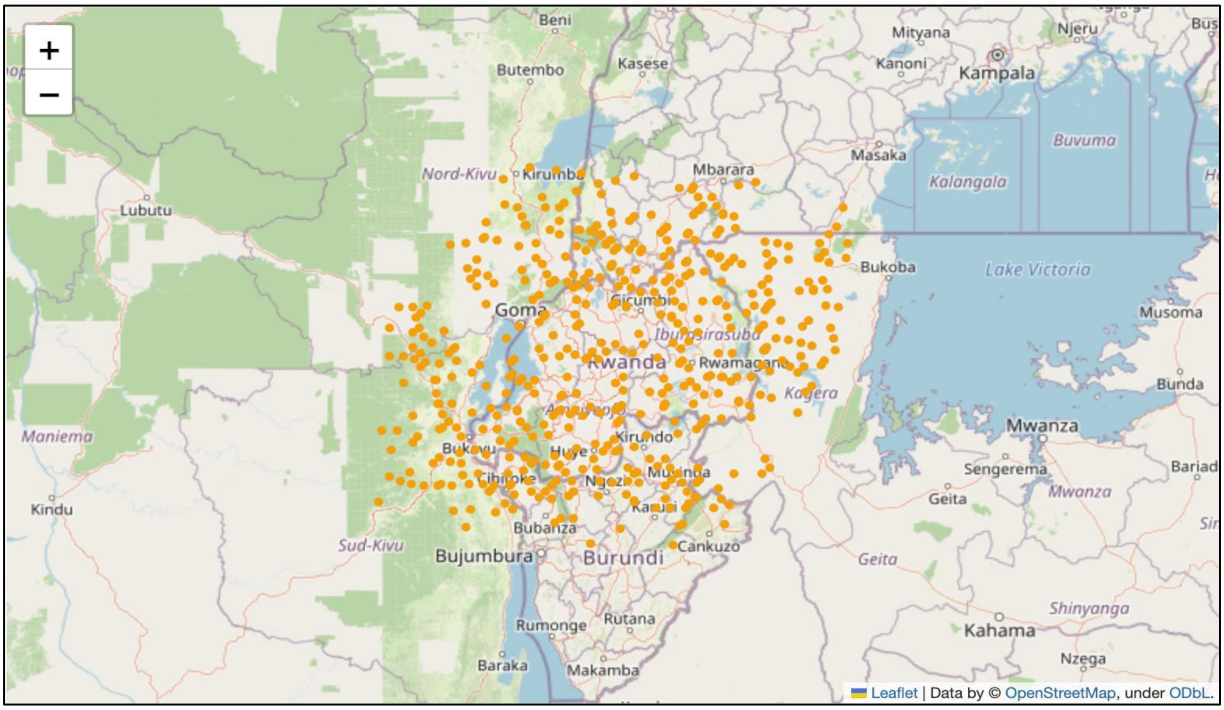


Figure 3.0: 497 unique geographical points where emissions were measured from

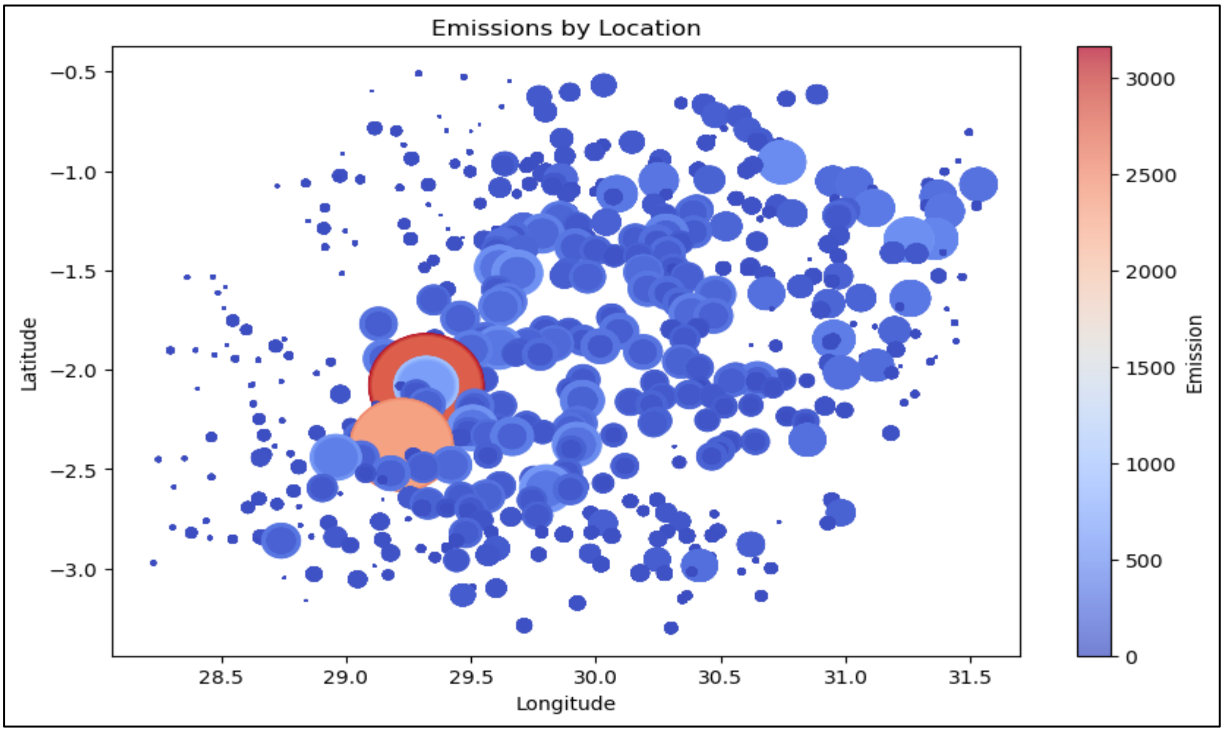


Figure 4.0: Emissions by Location (Longitude by Latitude)

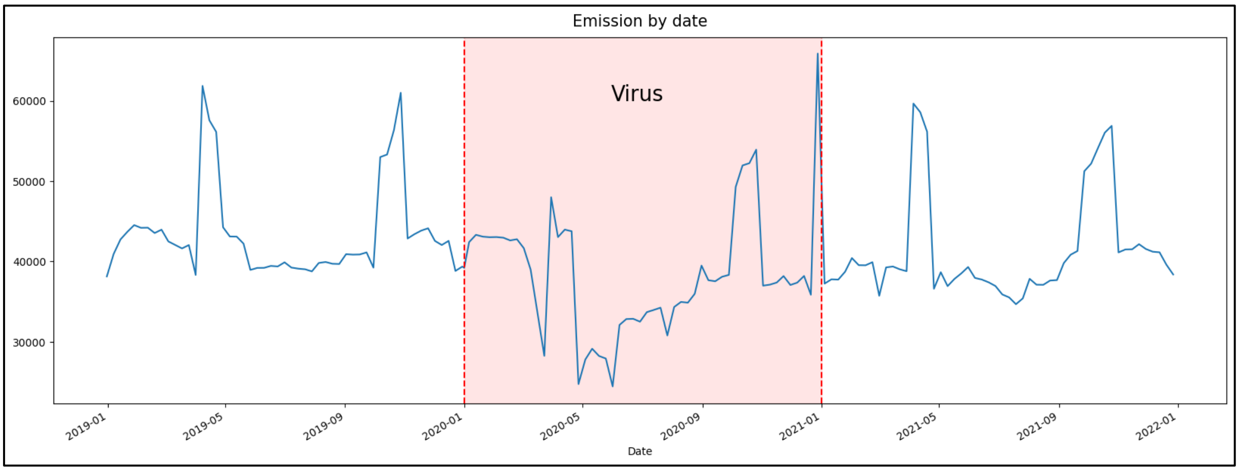


Figure 5.0: Emissions by Date highlighting the COVID-19-time range.

The time series chart indicates a significant impact of the COVID-19 pandemic on CO_2 emissions, particularly in the second quarter of 2020. This period is marked as an outlier due to its unique trends, influenced by global lockdowns and economic slowdowns. For modeling purposes, this anomaly could lead to overfitting if not addressed properly. In subsequent preprocessing steps, strategies must be implemented to mitigate this risk, ensuring that the model's predictive power is not compromised by these atypical data points.

Metrics:

For a time, series problem like CO_2 emission prediction, the choice of metrics is crucial to evaluate the model's performance accurately. Some of the metrics to be considered:

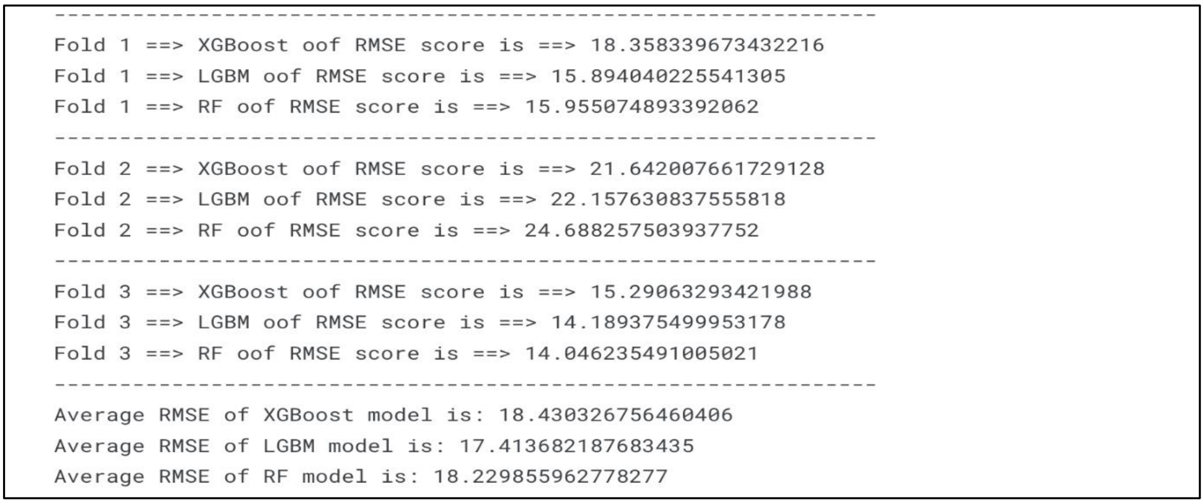
- **RMSE (Root Mean Square Error):** This metric is widely used because it penalizes larger errors more due to squaring the residuals. This makes it sensitive to outliers, which can be beneficial when large errors are particularly undesirable.
- **MAE (Mean Absolute Error):** MAE provides a straightforward average of absolute errors and is less sensitive to outliers compared to RMSE. It is useful when you want to avoid giving extra weight to the larger errors.
- **MAPE (Mean Absolute Percentage Error):** MAPE expresses accuracy as a percentage, which can be more intuitive. However, it can be problematic with values close to zero and can give infinite or undefined values for actual values of zero.

RMSE is chosen for this CO_2 emission prediction problem for several reasons:

- **Sensitivity to Large Errors:** RMSE is extremely sensitive to outliers, which means it gives a higher weight to large errors. This characteristic is particularly useful when predicting CO_2 emissions, as large deviations from actual emissions can have significant environmental impacts.
- **Consistency with Variance:** Since RMSE is in the same units as the predicted variable and squares the errors before averaging, it is consistent with the variance and standard deviation measures. This makes it easier to relate to the data's statistical properties.
- **Performance on Continuous Data:** RMSE is favored for continuous data, which is typical in time series problems like emission levels, as it can adequately measure the average magnitude of the error.
- **Model Selection:** When comparing various models, RMSE can help differentiate between models that might otherwise seem similar when assessed with less sensitive metrics like MAE.

Despite its advantages, RMSE is not without its drawbacks. It can be disproportionately influenced by large errors, which may not be representative of the model's performance if these errors are anomalies.

Modelling:



```
---------------------------------------------------------------
Fold 1 ==> XGBoost oof RMSE score is ==> 18.358339673432216
Fold 1 ==> LGBM oof RMSE score is ==> 15.894040225541305
Fold 1 ==> RF oof RMSE score is ==> 15.955074893392062
---------------------------------------------------------------
Fold 2 ==> XGBoost oof RMSE score is ==> 21.642007661729128
Fold 2 ==> LGBM oof RMSE score is ==> 22.157630837555818
Fold 2 ==> RF oof RMSE score is ==> 24.688257503937752
---------------------------------------------------------------
Fold 3 ==> XGBoost oof RMSE score is ==> 15.29063293421988
Fold 3 ==> LGBM oof RMSE score is ==> 14.189375499953178
Fold 3 ==> RF oof RMSE score is ==> 14.046235491005021
---------------------------------------------------------------
Average RMSE of XGBoost model is: 18.430326756460406
Average RMSE of LGBM model is: 17.413682187683435
Average RMSE of RF model is: 18.229855962778277
```

Figure 5.0: Key Statistics

- The LightGBM model consistently showed robust performance, leading with the lowest average RMSE, indicating it is potentially the most accurate and reliable model for predicting CO_2 emissions in this context.

- The XGBoost model, while exhibiting higher variability across folds, holds promise due to its competitive average RMSE, suggesting with fine-tuning, it could perform comparably.

- The Random Forest model, despite its higher average RMSE, could offer advantages in interpretability and robustness against overfitting, valuable for understanding feature importance and model behavior.

- The variability of RMSE scores across folds for all models suggests that feature selection, hyperparameter optimization, and ensemble methods could be critical to improve model stability and performance.

- These insights serve as a baseline for iterative model improvement and further exploration of data preprocessing and feature engineering to enhance predictive accuracy.

Conclusions

Advantages:

- Machine learning models, especially gradient boosting methods like LightGBM, demonstrated strong predictive capabilities.
- The use of time series data allowed for the identification of temporal trends, seasonality, and cyclic behavior in emission patterns.
- Model diversity with algorithms like XGBoost, Random Forest, and CatBoost offered a range of perspectives on data and helped in selecting the best-performing model.
- Predictive models can be instrumental in anticipating future emissions and setting benchmarks for environmental standards.
- The analysis provides valuable insights into the effects of global events, such as pandemics, on environmental indicators.
- It highlights the importance of considering external factors (like a pandemic) in predictive modeling.

Limitations:

- The global pandemic presented a unique challenge for models as they might not generalize well to periods with abrupt, unforeseen changes.

- The variance in model performance across different validation folds suggested data inconsistencies or the need for more robust feature engineering.
- Limitations in the collection of data during extraordinary events, COVID-19 in this case, lead to gaps in the dataset, affecting the model's accuracy.
- Potential biases may have arisen in the dataset given certain factors affecting emissions were not adequately recorded or considered.

Operational Recommendations

In our research, we arrived at the below operational recommendations:

- To further improve the model's resilience against anomalies, we would integrate outlier detection and handling techniques.
- For future research, enhanced data collection processes would need to be employed to reduce gaps in the dataset and to improve the quality of input data for future modeling efforts.
- Develop a system for real-time data analysis to promptly adjust to changes in emission patterns.
- Assembling models to blend the strengths of individual models, potentially increasing accuracy, and reliability.
- Apply scenario analysis using the models to understand the potential impact of various policy decisions or economic changes on CO_2 emissions.
- Foster an iterative approach to model development, including regular updates and refinements as more data becomes available or as circumstances change.
- Encourage collaboration with environmental experts to interpret model findings and incorporate domain knowledge into the predictive process.
- Promote transparency in the modeling process to build trust with stakeholders and facilitate the adoption of model recommendations in policymaking.