Predict future CO2 emissions in Rwanda

Group 8: Junzhe Wu, Ruining Tao, Yutao Zhou

1. Abstract

With the advancement of technology and the future of the Earth, the emission of carbon dioxide has become a concern for everyone. Africa is currently in a period of rapid development, which will bring a lot of carbon dioxide emissions. Therefore, effective monitoring, reasonable management and control, and prediction have become one of the methods to protect the Earth. Consequently, we utilize the combination of the Gaussian Mixture Model (GMM) and multiple regression models to predict future carbon dioxide emissions in Rwanda by analyzing the emissions of different substances. Compared with some other models such as XGBoost and LightGBM, Random Forest has the best performance. The project is based on Kaggle [1].

2. Introduction

The increasing level of carbon dioxide (CO2) emissions is a major contributor to climate change. Consequently, precise monitoring and prediction of CO2 emissions play a pivotal role in formulating effective mitigation strategies. Many parts of Africa need more comprehensive data, which impedes the development of targeted strategies for emission reduction in Africa. The primary objective of this project is to leverage data mining techniques to predict future CO2 emissions in Rwanda, a country in East Africa. This prediction will be based on open-source CO2 emissions data from Sentinel-5P [2] satellite observations. Through the integration of data mining techniques, we seek to empower stakeholders with valuable insights that inform decision-making processes and foster a more sustainable and resilient future for Rwanda.

3. Motivation

From a data mining perspective, this project offers a unique opportunity to apply and evaluate the efficacy of techniques such as clustering and regression algorithms in the context of predicting CO2 emissions. The complexity of environmental data presents a compelling real-world application for data mining methodologies. By addressing the challenges, this project aims to demonstrate the adaptability and utility of data mining in enhancing our understanding of environmental phenomena. Furthermore, from a real-world application standpoint, the accurate prediction of CO2 emissions in Rwanda has the potential to guide targeted policies, contributing to the global effort to combat climate change and promote sustainable development.

4. Related Work

3.1 Using Gaussian Mixture Model for Clustering on Climate Data

The Gaussian Mixture Model (GMM) assumes that the underlying data is generated from a mixture of Gaussian distributions. When the features in the dataset are close to a Gaussian distribution, it aligns well with the fundamental assumption of GMM. GMM has been widely used for prediction tasks on climate data. In Paçal et al.'s (2023) work of detecting the extreme temperature event, the authors found that the temperature data from the climate models can be

more accurately described using a mixture of multiple Gaussian distributions [3]. Ni et al. (2020) also utilized GMM to cluster streamflow data into multiple groups. Then each group was used to fit several single XGBoosts for streamflow forecasting tasks [4]. The author states that a cluster analysis-based model helps improve accuracy and capture complicated patterns.

3.2 Using XGBoost for Prediction Tasks on Climate Data

XGBoost, or eXtreme Gradient Boosting, is a powerful machine learning algorithm known for its effectiveness in various prediction tasks. It is particularly well-suited for time series data due to its ability to capture complex relationships and patterns. In the context of climate-related time series data, XGBoost has demonstrated notable success. For instance, Osman et al. (2021) applied XGBoost to predict groundwater levels in Malaysia [5]. Similarly, Ma et al. (2020) employed XGBoost to predict outdoor air temperature and humidity in China [6], and Fan et al. (2018) also utilized XGBoost for predicting daily global solar radiation in humid subtropical climates [7]. These examples highlight XGBoost's superiority over many baseline models in terms of accuracy, stability, and computational efficiency.

4. Methodology

4.1 Exploratory Data Analysis

1. Feature Analysis
   The training data contains 75 features. There are 4 index features (latitude, longitude, year, and week_no) and 1 target feature (CO2 emission). The remaining 70 features are the measurement information about various materials or indexes in the air collected by the Sentinel-5P [1] satellite. These 70 measurement features can be divided into 8 major groups: SulphurDioxide, CarbonMonoxide, NitrogenDioxide, Formaldehyde, UVAerosolIndex, Ozone, UvAerosolLayerHeight and Cloud. Each group contains a similar pattern of sub-features such as the density and angle information of the satellite. The training data contains 3-year information from 2019 to 2021, each has 53 weeks of info. The test data contain info for the year 2022, each has 49 weeks of info.
2. Target Value Analysis

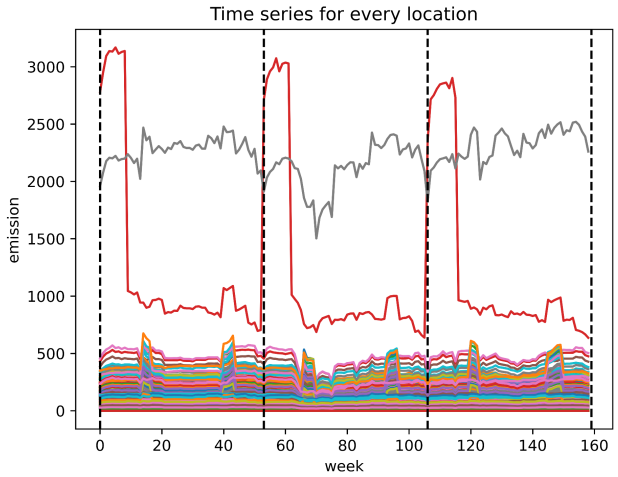


Figure 1. Time series representation of CO2 emission for every location

Figure 1 is the time series for CO2 emissions at each location. Among 497 unique locations, it is obvious that there are two locations (red and gray) with extremely high

emissions and they have special patterns; while all other locations look similar with relatively lower emissions. Also, there are 15 locations whose CO2 emissions are constantly zero. With further investigation, there exists seasonality in the CO2 emission, i.e. the yearly patterns (which repeat every 53 weeks).

4.2 Feature Engineering

1. Handling Missing Values
   After analyzing all 75 features, among 8 groups of features, UVAerosolIndex is the worst on data completeness with over 99% of training data and 91% of testing data missing, and there is a significant data drift between those features in training and test sets. Thus this feature sub-group is eliminated.
2. Fixing CO2 emission in the COVID-19 Period
   It is known that the whole world was experiencing a difficult time due to the COVID-19 pandemic. The COVID-19 also affected the CO2 emission in Rwanda. Figure 2 shows that, for every quarter of 2020 and 2021, how much the emissions have grown compared to the same quarter of the previous year. Every blue dot corresponds to a location. The x-axis is the average emission of the location, and the y-axis is how much the emission has increased or decreased. The lines in the diagrams are regression lines that serve to estimate the annual growth rate.

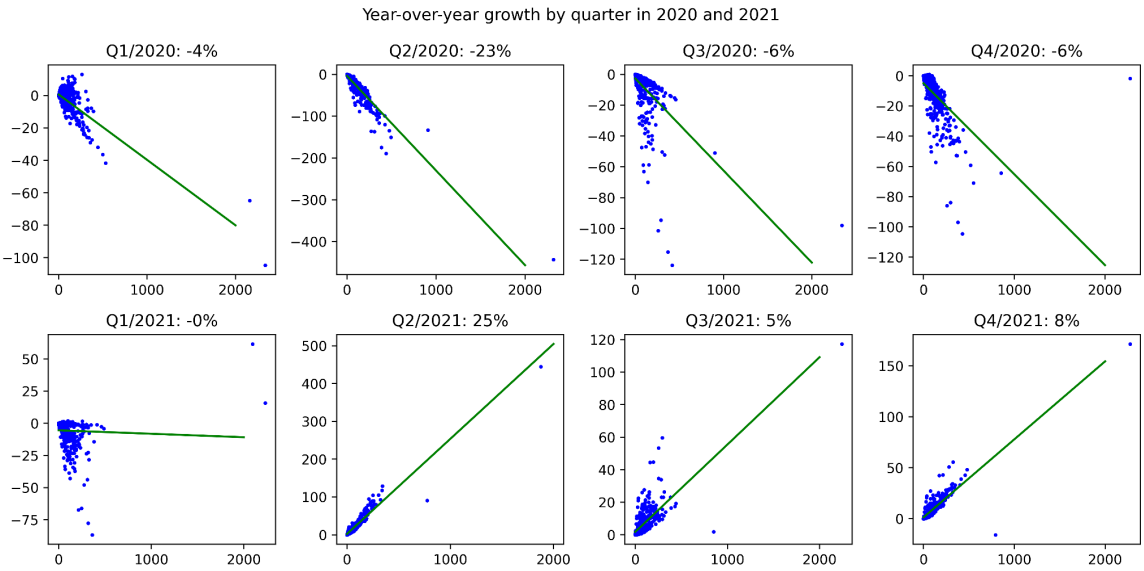
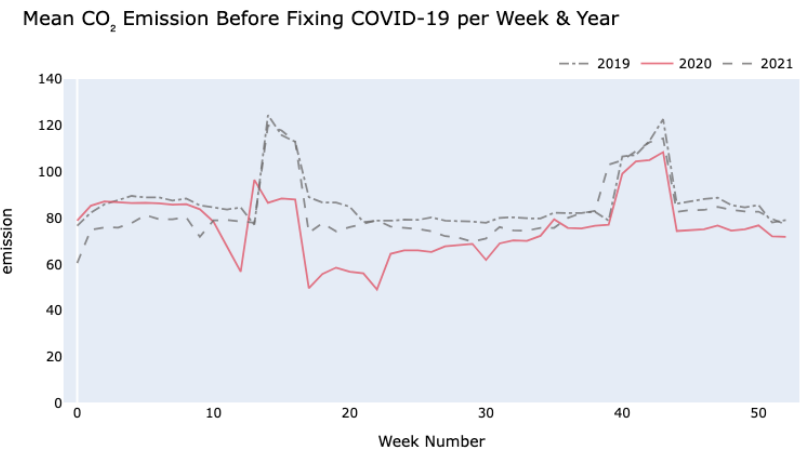


Figure 2. Year-over-year growth of CO2 emission by quarter in 2020 and 2021

It is very clear that COVID-19's influence on CO2 emission. Q1 of 2020 has shown a decreasing tendency. The most affected quarter is Q3 in 2020 with a more than 23% decrease in emissions. While the decreasing trend for Q3 and Q4 in 2020 is relatively small. When it comes to Q2 of 2021, the CO2 increased by 25% compared with Q2 of 2020, which could be treated as a recovery from the pandemic.
To address the potential negative influence on the model, such as over-fitting, due to the outlier, the COVID-19 period, the approach is to change 2020's emissions to the average level of 2019 and 2021 to eliminate the influence of irregular patterns. The difference before and after fixing the COVID-19 period is shown in the two diagrams below.



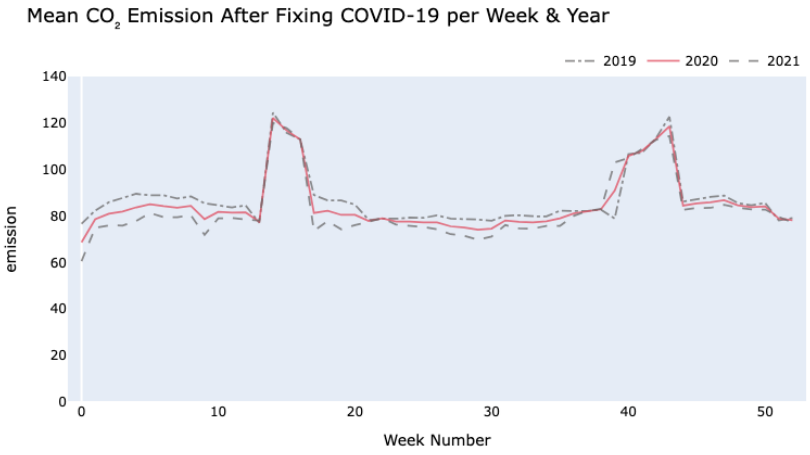


Figure 3 & 4. Mean CO2 Emissions Before and After Fixing CO2 Emission in COVID-19 Period

3. Remove Satellite Angle-related Features
   As described in *§4.1.1*, there are 8 groups of features, each containing a similar pattern of sub-features such as the density and angle information of the satellite. The angle-related features describe the position of the satellite when the data is measured. Based on common sense, the angle data is barely related to the actual CO2 emission but instead could add much noise to the whole dataset and create a negative influence on the model performance. Thus, all the angle-related features are removed.

4.3 K-means & GMM for Clustering

We choose Gaussian Mixture Models (GMMs) because GMMs can adapt to different cluster shapes, such as ellipses, and better reflect the distribution of variations found in real-world data. While K-means only works for similarly sized spherical clusters, GMM can effectively handle mixture distributions and provide density estimates. Although we chose the Gaussian Mixture Model (GMM) because of its advanced clustering capabilities, K-means can still play an important role in determining the optimal number of clusters for a data set. So, after a thoughtful discussion, we decided to use K-means to determine the number of clusters for the dataset and GMM to get good clustering.

4.4 Choose Regression Models

Since our response variable Carbon Dioxide emission is continuous. Thus, we should choose a regression model. We thought about linear regression, but since our data is very complex and not that correlated with emission. Thus, we decided to choose Random Forest and its related model. To be able to capture the non-linear relationship variables, random forest will have better performance than linear regression. Also, Random Forest includes built-in mechanisms (like feature randomness) to avoid overfitting. Thus, we choose XGBoost, LightGBM, and Random Forest itself. Each of them has its advantages.

5. Empirical Results

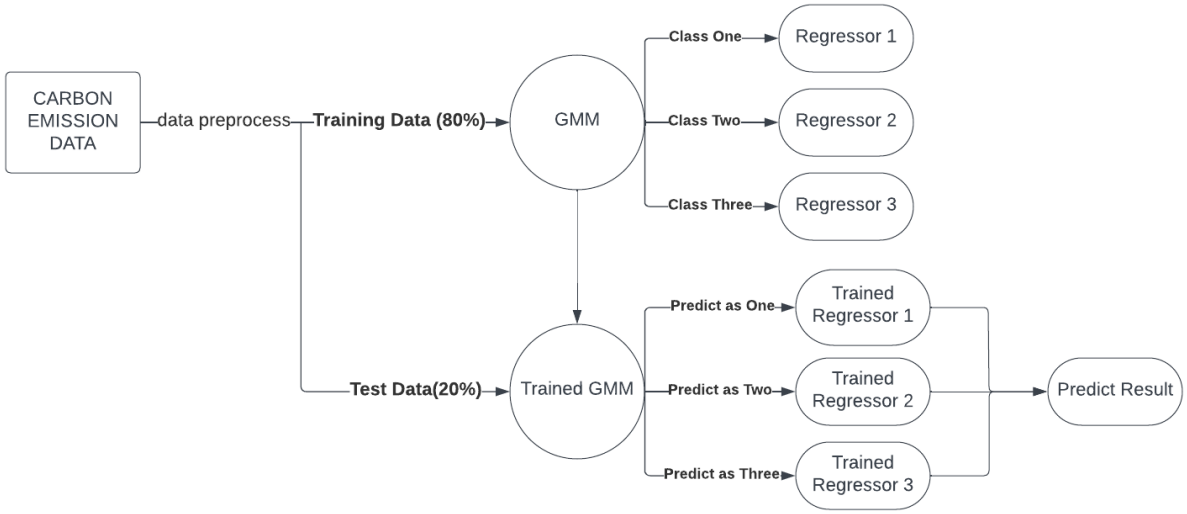


Figure 5. Diagram of Data Segmentation and Model Construction

There are two questions that we want to discover in this dataset. First, we want to train a model that can help the Rwanda government to predict CO2 better. Secondly, we want to explore whether the clustered data can capture some of the features that we were unable to obtain using the regression model. Thus we designed the model in the above Figure 5. There are two stages in our design in the model training process.

Firstly, clustering the data was essential, and the Gaussian Mixture Model (GMM) was selected over K-means due to its enhanced capabilities. The dataset's unique geographical and Timeliness characteristics prompted the use of GPS and CO2 emission information to classify the data. With the best number of clusters undetermined, we explored a range from 1 to 12 clusters using K-means to establish the most effective clustering number. The result is shown in Figure 6. Based on this information, we decided to cluster 3 classes on our data.

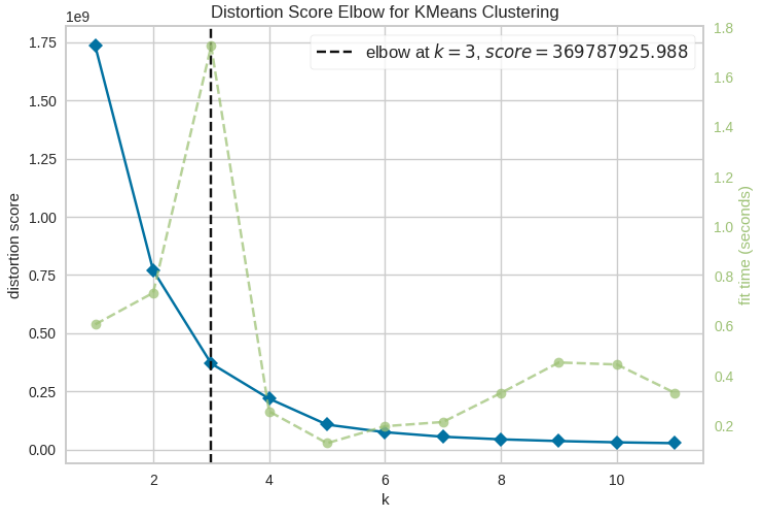


Figure 6. Distortion Score Elbow for K-Means Clustering

In the second stage, we trained each data class and set aside a test dataset, as no ground truth for carbon emissions was available in the test data provided. Consequently, we allocated 20% of the data from each class as the test set. This resulted in 80% of the data forming the training set, with the remaining 20% combined and shuffled to create the test set. After data segmentation, we trained on four different regression models: LightGBM, XGBoost, and Random Forest. To prove the benefit of using GMM for pre-training data classification, we trained XGBoost on unclustered data as a baseline model.

Given the continuous nature of carbon emission data, a direct comparison of estimated results is not feasible. Therefore, we employed Mean Absolute Error (MAE) as the metric to assess our model's performance.

Model Name	MAE for class 1		MAE for class 2		MAE for class 3		MAE for all data	
	Train	Valid	Train	Valid	Train	Valid	Train	Valid
Baseline	x	x	x	x	x	x	42.695	71.790
GMM + lightGBM	14.50	15.64	35.14	141.41	63.15	69.88	33.572	38.619
GMM + XGBoost	11.52	15.63	0.0003	134.65	50.54	71.96	26.358	38.438
GMM + RF	5.970	15.94	41.92	146.14	27.09	70.67	**14.569**	**38.300**

Table 1. Experiment Results for Multiple Coupling Models

The results in Table 1 indicate that the combination of Gaussian Mixture Models (GMM) with Random Forest (RF) yields the most favorable outcomes, as evidenced by the lowest mean absolute error across both training and validation datasets. This finding aligns with our objectives, demonstrating that clustering data and subsequently training a regressor for each class significantly enhances model performance. Compared to the baseline model, this approach has halved the mean absolute error, marking a great improvement. A notable limitation, however, is the elevated train and validation error for class 2. All three methods underperformed for class 2, which can be attributed to its notably smaller sample size. Class 2 primarily comprises data from locations with exceptionally high carbon dioxide emissions, and the limited data points in this class pose challenges in achieving better accuracy. Despite this, the overall performance remains satisfactory.

6. Conclusion

In this project, we conduct a detailed analysis and processing of the CO2 emission data in Rwanda. Then utilizing GMM and multiple regression models to predict future CO2 emissions. The results turned out that coupling GMM with Random Forest is the best among all models. For the future direction, we might need more data and find more related variables. Also, trying another model such as the Hidden Markov Model or some advanced machine learning models may have potentially good performance.

Reference:

[1] https://www.kaggle.com/competitions/playground-series-s3e20/data
[2] https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-5p
[3] Paçal, A., Hassler, B., Weigel, K., Kurnaz, M. L., Wehner, M. F., & Eyring, V. (2023). Detecting extreme temperature events using Gaussian mixture models. Journal of Geophysical Research: Atmospheres, 128.
[4] Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J., & Liu, J. (Year). Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. Journal of Hydrology, 586.
[5] Osman, M., et al. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor, Malaysia. Journal Name, Volume(Issue), Page Range.
[6] Xiaoming Ma, C. Fang, J. Ji (2020). Prediction of outdoor air temperature and humidity using Xgboost. Journal Name, Volume(Issue), Page Range.
[7] Fan, X., et al. (2018). Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. Journal Name, Volume(Issue), Page Range.