DLI 2025 COMMUNITY CHALLENGE

Title: Democratizing CO_2 Monitoring in Nigeria with AI and Low-Cost Local Sensors

1. Introduction

Overview of the Problem and Its Significance

Air pollution has emerged as one of the most pervasive environmental challenges of the 21st century, posing a serious and escalating threat to human health, ecological balance, and the global climate system. It involves the introduction of a wide array of harmful and toxic substances into the atmosphere, including but not limited to particulate matter (PM), carbon dioxide (CO_2), carbon monoxide (CO), nitrogen oxides (NOx), sulphur dioxide (SO_2), and volatile organic compounds (VOCs). These pollutants, once emitted, can interact with each other and with natural environmental elements to produce secondary pollutants like smog and ground-level ozone. The consequences are far-reaching, ranging from reduced agricultural productivity, degraded air and water quality, biodiversity loss, and altered climate patterns, to direct and often severe impacts on human health such as chronic respiratory conditions, cardiovascular diseases, neurological disorders, and increased mortality.

In countries like Nigeria, the problem of air pollution is particularly pronounced and multifaceted. Rapid and largely unregulated urban expansion, coupled with population growth, industrialization, and widespread reliance on fossil fuels for transportation, power generation, and household energy use, has led to significant increases in atmospheric pollutant concentrations. The burning of biomass, use of generators, vehicle exhaust emissions, and industrial flaring are major contributors. Despite growing awareness, efforts to monitor and manage air quality have been hampered by systemic gaps in infrastructure, limited financial and technical resources, and a general absence of reliable, real-time environmental data.

Among all air pollutants, carbon dioxide (CO_2) occupies a central role in climate dynamics. As a key greenhouse gas, CO_2 traps heat in the atmosphere through the greenhouse effect, contributing to rising global temperatures, sea level rise, and increased frequency and intensity of extreme weather events. Recent climate science research shows that nearly one-third of all global CO_2 emissions since the industrial era began have occurred in just the past two decades, a period marked by intensified economic activity and energy consumption. While industrialized countries remain the largest historical emitters, emerging economies like Nigeria are increasingly contributing to global CO_2 levels through the combustion of fossil fuels used in power plants, vehicular engines, and manufacturing industries. Unfortunately, developing nations often face greater difficulty in adapting to the consequences of climate change due to fragile health systems, economic volatility, and infrastructural deficits.

Monitoring CO_2 emissions accurately and at scale is essential to guide policy decisions, enforce environmental standards, and raise public awareness. However, high-accuracy reference-grade CO_2 monitoring equipment, including non-dispersive infrared analyzers and advanced spectrometers, are prohibitively expensive, require frequent calibration, and are largely unavailable across many parts of Africa. This situation gives rise to what is often described as “data darkness”, the lack of accurate, continuous, and geospatially granular CO_2 data, particularly in underserved and rural regions. In the absence of such data, decision-makers, scientists, and local communities are left with blind spots that hinder effective climate action, health interventions, and infrastructure planning.

To bridge this critical gap, our study proposes a cost-effective and scalable machine learning (ML)-driven solution that utilizes low-cost, Nigerian-manufactured analog gas sensors to generate reliable estimates of ambient CO_2 concentration. These sensors, developed by local innovation hubs such as Chemotronix, are affordable and widely deployable but inherently noisy and susceptible to sensor drift, cross-sensitivity, and environmental variability. The core objective of our approach is to leverage machine learning algorithms to transform raw, uncalibrated analog sensor data into accurate, standardized CO_2 measurements by mapping them to values obtained from high-fidelity reference sensors. This software-based calibration framework enables even resource-constrained regions to gain access to real-time, near-reference quality CO_2 data without incurring the cost of professional-grade instrumentation.

Our model is designed not only to perform well in terms of predictive accuracy but also to maintain robust performance across different sensor units, temporal scales, and diverse environmental contexts, including varied temperature and humidity levels, both indoors and outdoors. Moreover, our system is able to simulate the conditions encountered in real-world deployment scenarios, enabling rapid adaptation to new sensor installations without the need for extensive reconfiguration.

Importantly, this initiative is aligned with broader global goals aimed at enhancing climate resilience, promoting environmental justice, and ensuring sustainable development. Recent empirical research has highlighted the significant influence of extreme weather conditions such as heatwaves and cold snaps on CO_2 emissions, with studies showing that CO_2 output during such events can surge by up to 27% due to increased energy demand for heating and cooling. This underscores the importance of high temporal resolution CO_2 monitoring systems, which are crucial for understanding short-term emission fluctuations, anticipating peak pollution periods, and developing real-time mitigation strategies. By enabling affordable, accurate, and continuous monitoring, our system offers a transformative step toward inclusive environmental intelligence and informed climate governance in Africa and beyond.

2. Methodology

AI Approach

In this study, we employed a supervised machine learning strategy aimed at solving a multivariate regression problem: predicting ambient carbon dioxide (CO_2) concentration levels using sensor readings obtained from locally manufactured, low-cost gas-sensing IoT devices. The primary objective of our modeling pipeline was to build robust and highly generalizable predictive models capable of learning the latent relationships between noisy analog sensor outputs and the corresponding accurate CO_2 measurements obtained from reference-grade instruments. These analog outputs are inherently unstable due to environmental fluctuations and sensor degradation over time. Therefore, the models needed to be not only accurate but also resilient across different operating conditions and sensor units.

To tackle this, we explored and implemented a family of tree-based ensemble regression models methods that are particularly well-suited to the types of data challenges we faced. These included non-linearity, feature collinearity, sensor drift, and moderate-to-high levels of noise. Ensemble models such as Random Forest, Extra Trees, and gradient boosting variants are known for their ability to reduce variance, increase prediction stability, and handle heterogeneous feature interactions. Additionally, these models offer feature importance metrics and interpretability tools, which are critical in environmental monitoring applications where transparency and trustworthiness are essential. Their scalability and robustness made them ideal candidates for our use case, where real-time, low-latency inference is important and data quality varies significantly.

Dataset Used and Data Collection Process

The dataset utilized for this project was made available through the Zindi Africa data science competition platform. The competition, which aimed to inspire Africa-centric technological solutions to environmental problems, specifically focused on the use of machine learning to enhance CO_2 emission prediction from low-cost sensors. The dataset comprises a total of 7,307 labeled instances, each consisting of synchronized readings from gas and environmental sensors, along with a target CO_2 concentration value measured in parts per million (ppm).

The sensor data were collected from three distinct Nigerian-manufactured IoT sensor prototypes designated as Alpha, Beta, and Charlie produced by Chemotronix, a hardware innovation company based in Nigeria. These devices integrate analog gas sensors such as MQ7, MQ9, MG811, and MQ135, along with temperature and humidity sensors. Data acquisition took place under a diverse set of environmental conditions that spanned indoor and outdoor settings across various Nigerian climate zones, including humid coastal areas, dry savannah regions, and dense urban corridors.

A defining feature of the dataset is its inclusion of ground-truth CO_2 measurements obtained from high-accuracy, professionally calibrated reference sensors**.** These measurements serve as the labels for the supervised learning task. As such, the entire modeling process essentially constitutes a calibration simulation, whereby our machine learning models learn to map unprocessed, noisy analog readings to the precise CO_2 concentrations provided by reference instrumentation. This mapping represents a software-based calibration layer that enhances the practical utility of low-cost sensors, enabling them to deliver near-reference quality predictions at a fraction of the cost.

The structure and variability embedded within the dataset also support the development of models that generalize across different sensor builds and environmental contexts. By learning from multi-device data collected over time and space, our models gain the capacity to adapt to new deployments and to sustain performance under realistic operating conditions an essential capability for scalable air quality monitoring in resource-constrained environments.

`Sensor	**Target Gases**	**Typical Applications**
MQ7	Carbon Monoxide (CO)	Home and industrial CO detection (e.g., from heaters, engines, stoves)
MQ9	Carbon Monoxide (CO), Methane (CH_4), Liquefied Petroleum Gas (LPG)	Fire alarms, gas leak detectors, industrial safety
MG811	Carbon Dioxide (CO_2)	Indoor air quality, HVAC systems, greenhouses
MQ135	Ammonia (NH_3), Benzene (C_6H_6), Carbon Dioxide (CO_2), Smoke	General air pollution monitoring, smart air purifiers, ventilation control systems
Temperature and Humidity	Atmospheric Temperature and Relative Humidity	Weather and environmental monitoring
Ground-truth CO_2	Calibrated CO_2 concentration (ppm) used as the prediction target	Provided as ground truth for supervised learning

Data Pre-processing

To ensure that the dataset was properly structured, cleaned, and optimized for training robust machine learning models, we undertook a series of carefully considered pre-processing steps. These steps were designed to address common issues associated with raw sensor data, such as noise, skewed distributions, variable scales, and potential outliers while preserving the underlying patterns critical for predictive accuracy.

- Missing Values Check: A thorough examination was conducted across all feature columns to identify any missing or null values. Fortunately, the dataset was found to be complete, with no missing entries, which allowed us to proceed without the need for imputation or data exclusion.
- Outlier Management Using Robust Scaling: Although outliers were present particularly in analog sensor readings due to environmental noise or sudden emission

spikes, they were not discarded, as they may represent genuine phenomena. Instead, we applied RobustScaler, a scaling technique that uses the median and interquartile range (IQR) to rescale feature values. This method is more robust to the presence of extreme values compared to traditional Min-Max or StandardScaler approaches, which are sensitive to outliers. By applying RobustScaler, we retained the influence of significant data points while mitigating their disproportionate effect on the learning algorithm.

- Logarithmic Transformation for Skewness Reduction: Several sensor features, such as gas concentration analog outputs, exhibited highly skewed distributions. To normalize these distributions and stabilize the variance across the dataset, we performed a log transformation on selected features. This transformation compressed the scale of extreme values, helping the model to learn more effectively from the data by reducing heteroscedasticity and enhancing symmetry.
- Feature Engineering via Interaction Terms: Beyond basic pre-processing, we introduced domain-specific interaction features by multiplying sensor values that are known to co-respond to overlapping gas profiles. For example, the product of MQ7 × MQ9 was computed to capture potential interactions between CO and methane sensing behaviour, while MG811 × MQ135 highlighted the joint sensitivity of CO_2-specific and multi-gas sensors. These engineered features enriched the model's input space with non-linear patterns and cross-sensor dynamics that might not be evident through individual sensor signals alone.

This comprehensive pre-processing pipeline laid a strong foundation for effective learning and ensured that the machine learning models were trained on a dataset that was well-conditioned, informative, and representative of real-world sensor behaviour.

Model Selection and Justification

To develop a high-performing predictive system for estimating ambient CO_2 concentrations from analog sensor data, we trained and evaluated five advanced ensemble-based machine learning algorithms. These models were chosen based on their proven success in structured tabular data problems, particularly in noisy environments with complex feature interactions characteristics that closely match our dataset.

- Random Forest Regressor: This is a bagging ensemble technique that constructs multiple decision trees using bootstrapped subsets of the training data and averages their predictions. Its robustness, interpretability, and minimal need for hyperparameter tuning make it a strong baseline model. It excels at handling non-linear relationships and offers built-in feature importance metrics.
- Extra Trees Regressor: An extension of Random Forest, Extra Trees introduces additional randomness during tree construction by selecting cut-points at random rather than using the optimal split. This added randomness increases model diversity, reduces variance, and often improves generalization particularly useful for datasets like ours that contain noise and potential sensor drift.
- XGBoost Regressor: XGBoost (Extreme Gradient Boosting) is a highly optimized implementation of gradient boosting decision trees. It is well-regarded for its scalability, efficient handling of missing values, and built-in regularization, which reduces the risk of overfitting. XGBoost performs exceptionally well in ML competitions and was included due to its ability to capture complex, non-linear signal patterns in the data.

- LightGBM: Light Gradient Boosting Machine (LightGBM) is another gradient boosting algorithm designed to handle large-scale data efficiently. It is optimized for speed and memory usage through histogram-based decision trees and supports parallel and GPU learning. LightGBM is particularly suited for high-dimensional datasets and tends to converge faster than traditional GBDT implementations.
- CatBoost: This model is specifically designed to handle categorical variables without extensive preprocessing, thanks to its unique approach to encoding. It is more resilient to noisy data, multicollinearity, and small datasets, making it an excellent fit for real-world sensor applications where raw data is often imperfect.

These models were selected based on the following key criteria:

- High predictive accuracy on small- to medium-sized tabular datasets
- Ability to resist overfitting in the presence of sensor noise, drift, or cross-sensitivity
- Capability to interpret and visualize feature importance, aiding explainability
- Scalability for real-time inference and deployment in edge or cloud systems

Hyperparameter Optimization Using Optuna

To maximize the performance of each model, we used Optuna, a state-of-the-art hyperparameter optimization framework that leverages Bayesian optimization principles to explore and refine the search space efficiently. Unlike traditional grid search or random search methods, Optuna builds a probabilistic model of the objective function (e.g. RMSE) and uses it to intelligently suggest hyperparameter combinations that are more likely to improve performance.

The tuning process in Optuna involves three main iterative steps:

1. Suggest: The system proposes a new hyperparameter configuration based on the historical performance of past trials.
2. Evaluate: The machine learning model is trained and evaluated using the suggested configuration. Performance metrics such as RMSE are computed.
3. Update: The performance feedback is used to update the internal model, narrowing the search space and improving future hyperparameter suggestions.

Optuna’s advanced capabilities such as early stopping, parallel trial execution, pruning of underperforming trials, and adaptive refinement of the search space significantly improved the efficiency and effectiveness of the tuning process. This was especially valuable for more complex models like XGBoost, LightGBM, and CatBoost, where the interplay between hyperparameters can be non-linear and computationally expensive to explore exhaustively.

By employing this intelligent tuning framework, we were able to achieve faster convergence, more stable results, and better generalization performance without incurring the high computational cost typically associated with exhaustive search techniques. This combination of powerful models and efficient tuning proved critical in delivering a high-performing, scalable predictive system for air quality monitoring using low-cost IoT sensors.

Evaluation Strategy

To ensure that our machine learning models were not only accurate during training but also resilient, adaptable, and generalizable in real-world deployment scenarios, we implemented a robust and comprehensive evaluation strategy. This multi-pronged approach included rigorous validation techniques, statistical performance metrics, device-aware evaluation protocols, and interpretability mechanisms aimed at boosting transparency and model trustworthiness.

Cross Validation (cv) We applied a 5-fold cross-validation strategy to assess the consistency and generalizability of the model's predictions. The dataset was systematically divided into 5 equally sized folds. For each of the five iterations, the model was trained on four of these folds and tested on the one that was held out, ensuring that every data point was eventually used for both training and testing.

The benefits of this strategy include:

- Comprehensive data utilization: Every observation contributes to both model training and validation, thus reducing the risk of biased performance estimates.
- Reduced performance variance: Cross-validation minimizes overfitting risks and ensures that the model's accuracy is not dependent on a specific subset of the data.
- Reliable metric aggregation: By averaging results over multiple iterations, we obtain stable performance indicators that reflect the model's robustness.

Evaluation Metrics We evaluated model performance using three complementary regression metrics:

- Root Mean Squared Error (RMSE): Quantifies the square root of the average squared difference between predicted and actual values, making it sensitive to large prediction errors.
- Mean Absolute Error (MAE): Represents the mean of the absolute differences between predicted and actual values, providing a clear, interpretable measure of average prediction error.
- R^2 Score: Measures the proportion of variance in the target variable that is explained by the model. An R^2 value closer to 1 indicates a strong predictive relationship.

These metrics collectively provided a multi-faceted view of the model's predictive accuracy, error tolerance, and explanatory power.

Device-Aware Validation

Since the dataset included measurements from three distinct IoT sensor devices Alpha, Beta, and Charlie we adopted a device-aware cross-validation framework. This method ensured that models were tested on sensor data from hardware not seen during training.

Implementation steps included:

- Rotating which device's data served as the validation set, while training the model on the remaining two.
- Repeating the rotation until all devices had served as the hold-out set.

This approach mimics deployment scenarios and:

- Validates generalizability across heterogeneous hardware environments.
- Mitigates overfitting to specific sensor units or conditions.
- Improves deployment readiness by simulating unseen operational data from new or evolving sensor configurations.

Model Explainability Using SHAP Understanding how input features influence predictions is crucial for deploying AI in environmental science. To this end, we integrated SHAP (SHapley Additive exPlanations), a game-theoretic tool for post-hoc model interpretability.

SHAP assigns each feature a contribution score its Shapley value reflecting how it influences a model's prediction for each instance. This technique allowed us to:

- Quantify feature contributions at both global and local levels.
- Enhance transparency in decision-making for non-technical stakeholders.
- Identify unexpected behaviors or sensor malfunctions based on SHAP anomalies.

SHAP Analysis Key Findings:

- MG811, the dedicated CO_2 sensor, emerged as the most important feature, confirming its predictive relevance.
- MQ135, sensitive to a mix of gases including CO_2, added auxiliary signal strength, especially in complex pollution settings.
- Temperature and humidity, though indirect, influenced model behavior due to their effects on sensor sensitivity and gas behavior.
- Engineered interactions such as MG811 × MQ135 and MQ7 × MQ9 contributed significantly to capturing nuanced, non-linear gas relationships.

Deployment Stack To ensure that this system delivers value in practical applications, we built a real-time, interactive web-based platform using open-source tools. This web application is designed for end-users including researchers, public health officials, and educators to access insights easily and act on them.

Technology Stack:

- Python (Scikit-learn, XGBoost, CatBoost): For model training, evaluation, and saving predictive artifacts.
- Streamlit: For quickly building a user-friendly front-end that supports interactive predictions.
- Pandas, Seaborn, Matplotlib: For statistical plotting and in-app data visualizations like histograms, correlation matrices, and SHAP value summaries.
- GitHub: For version control, model reproducibility, and open collaboration.
- Streamlit Cloud: For hosting the app, enabling access from any device via browser.

User Functionality:

- Upload and analyze sensor data for instant CO_2 prediction.
- View sensor readings and prediction outcomes through interactive graphs.

- Examine SHAP-based interpretability plots to understand the reasoning behind each prediction.

By combining advanced modeling, explainability tools, and a streamlined deployment interface, we have created a reliable, transparent, and user-friendly tool that democratizes CO_2 monitoring and promotes sustainable innovation in environmental sensing.

3. Results and Findings

Exploratory Data Analysis (EDA)

To construct a thorough, nuanced, and statistically sound understanding of the dataset that would effectively support the development of a robust and scalable predictive model, a comprehensive and multifaceted exploratory data analysis (EDA) phase was undertaken. This stage of the project was critical in uncovering underlying trends, identifying significant patterns, detecting anomalies, understanding the variability within the dataset, and examining the interdependencies between sensor outputs and environmental factors such as temperature and humidity.

Temperature, being one of the key contextual environmental variables in the dataset, displayed an average recorded value of approximately 29.7°C. This central tendency aligns with the expected thermal profile of Nigeria, characterized predominantly by its tropical equatorial climate. The associated standard deviation of around 2°C indicates a moderate level of fluctuation across the entire set of records. The observed temperature values fell within a broad range, from a minimum of 23.44°C to a peak of 33.85°C. The interquartile range (IQR), which extended from 28.49°C to 31.72°C, revealed that half of the temperature observations were tightly clustered within this span, indicating a relatively stable thermal environment for the majority of recordings, punctuated occasionally by spikes likely driven by heatwave conditions, seasonal transitions, or microclimatic effects related to urban heat islands or device placement.

Humidity, another influential environmental condition with the potential to impact sensor performance and gas diffusion, demonstrated a slightly right-skewed distribution. This skewness reflects the occasional emergence of very high humidity conditions, which are not uncommon in tropical regions, particularly during the rainy season. The concentration of values around a moderately high central value, accompanied by frequent appearances of high-humidity extremes, underscores the importance of normalization techniques to handle these long-tail distributions and to ensure that modeling efforts remain robust against such variabilities.

Among the analog gas sensor readings, MQ7_analog which is primarily used for detecting carbon monoxide (CO) registered a mean reading of 4230.8, with a standard deviation of 1085, highlighting significant spread within the data. The values spanned a wide spectrum, from a minimum of 2380 to a high of 10379.5, with an IQR from 3362.5 to 5286.2. These findings suggest an operationally responsive sensor that captures a diverse range of CO events or interferences, with potential spikes attributable to high pollution levels or noise due to sensor cross-sensitivity.

Similarly, MQ9_analog, designed to sense both CO and methane (CH_4), recorded a mean value of 3976.7 and a standard deviation of 1098.7. The sensor outputs ranged from 1098.5 to

7919, and its IQR extended from 3181 to 4731, indicating that while most of the readings clustered within a predictable band, some lower-bound values may indicate either signal anomalies, atmospheric dilution, or low gas presence in cleaner microenvironments.

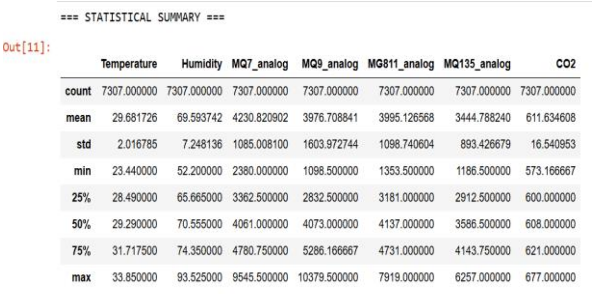
The MG811_analog sensor, the dedicated CO_2 sensing device in the setup, delivered an average value of 3995.1 with a standard deviation of 893.4. With a value range spanning from 1353.5 to 6257 and an IQR between 3181 and 4713.75, the sensor showed reliable responsiveness across a diverse set of ambient CO_2 levels, capturing both background and potentially elevated emission scenarios.

MQ135_analog, known for its multi-gas detection capability including ammonia (NH_3), nitrogen oxides (NOx), and CO_2, produced a mean analog output of 3444.8 with a standard deviation of 915.4. Its readings extended from a low of 1186.5 to a high of 6777, with an IQR ranging from 2912.5 to 4143.75. The breadth of detection validates its use as a broad-spectrum air quality sensor, especially in environments where gas mixtures are common.

For the target variable CO_2 concentration, measured in parts per million (ppm), the dataset showed a mean of 611.6 and a standard deviation of 16.5, suggesting less volatility compared to the analog sensor readings. The values ranged from 573.17 to 677 ppm, while the IQR extended from 600 to 621 ppm. This relatively narrow dispersion is consistent with atmospheric conditions, yet the presence of 310 unique values within this band signifies adequate granularity for regression modeling.

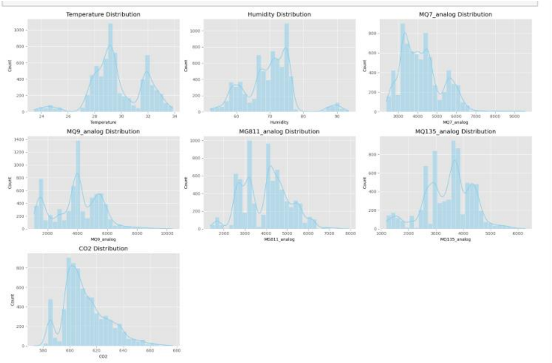
Metadata revealed three distinct sensing devices: Alpha, Beta, and Charlie, which were embedded as categorical identifiers and were treated accordingly during preprocessing to account for inter-device variation. The ID column, being row-unique and non-informative for prediction, was excluded from model training.

EDA visualizations including histograms, KDE plots, and boxplots revealed that most sensor variables exhibited either skewed or multimodal distributions. These visual analyses also exposed significant outliers, which reinforced the necessity for advanced scaling techniques such as RobustScaler and the deployment of tree-based ensemble models that are tolerant to distributional irregularities.



```
=== STATISTICAL SUMMARY ===
Out[11]:
```

	Temperature	Humidity	MQ7_analog	MQ9_analog	MG811_analog	MQ135_analog	CO2
count	7307.000000	7307.000000	7307.000000	7307.000000	7307.000000	7307.000000	7307.000000
mean	29.681726	69.593742	4230.820902	3976.708841	3995.126568	3444.788240	611.634608
std	2.016785	7.248136	1085.008100	1603.972744	1098.740604	893.426679	16.540953
min	23.440000	52.200000	2380.000000	1098.500000	1353.500000	1186.500000	573.166667
25%	28.490000	65.665000	3362.500000	2832.500000	3181.000000	2912.500000	600.000000
50%	29.290000	70.555000	4061.000000	4073.000000	4137.000000	3586.500000	608.000000
75%	31.717500	74.350000	4780.750000	5286.166667	4731.000000	4143.750000	621.000000
max	33.850000	93.525000	9545.500000	10379.500000	7919.000000	6257.000000	677.000000



Descriptive Statistics of our Dataset

Histogram Plot of each numerical features

Pearson Correlation Coefficients with CO_2 Target:

- MG811_analog: 0.1015
- MQ7_analog: 0.0937
- MQ135_analog: 0.0273
- MQ9_analog: -0.0264
- Humidity: -0.0324
- Temperature: -0.0504

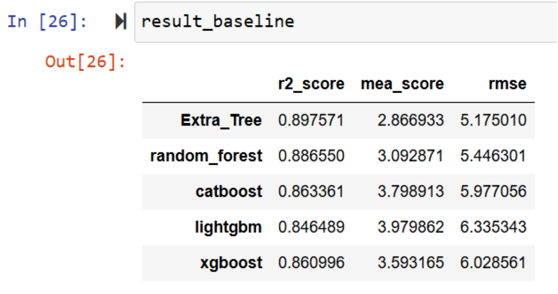
Key Observations and Interpretations:

- MG811 had the strongest direct linear association with the CO_2 target variable, reaffirming its central role in the predictive model.
- MQ7 followed closely with a modest positive correlation, indicating possible sensitivity to environmental conditions correlating with CO_2 presence.
- MQ9 and temperature displayed weak but consistent inverse correlations.
- Humidity displayed minor and inconsistent correlation, suggesting a possible indirect or non-linear influence.

Significance Testing and Non-Linear Correlations:

- Correlation coefficients for MG811, MQ7, and MQ135 were found to be statistically significant at standard p-value thresholds.
- Supplementary use of Spearman's rank correlation confirmed the existence of monotonic trends, supporting the use of interaction and non-linear terms in the modeling phase.

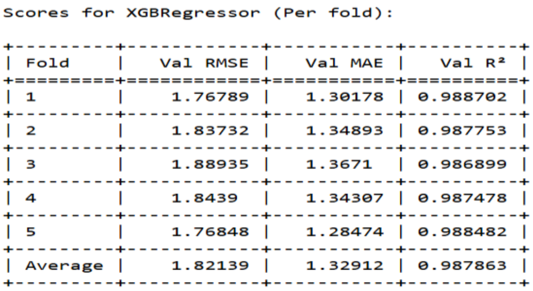
These analytical outcomes laid the foundation for thoughtful feature engineering, informed model selection, and provided insights into the importance of multi-sensor fusion and temporal dynamics in predicting CO_2 levels under varying conditions.



```
In [26]: result_baseline
Out[26]:
```

	r2_score	mea_score	rmse
Extra_Tree	0.897571	2.866933	5.175010
random_forest	0.886550	3.092871	5.446301
catboost	0.863361	3.798913	5.977056
lightgbm	0.846489	3.979862	6.335343
xgboost	0.860996	3.593165	6.028561

Figure 1 Baseline Models Evaluation



```
Scores for XGBRegressor (Per fold):

+----------+-------------+-------------+-----------+
| Fold     |    Val RMSE |     Val MAE |    Val R² |
+==========+=============+=============+===========+
| 1        |     1.76789 |     1.30178 | 0.988702  |
+----------+-------------+-------------+-----------+
| 2        |     1.83732 |     1.34893 | 0.987753  |
+----------+-------------+-------------+-----------+
| 3        |     1.88935 |     1.3671  | 0.986899  |
+----------+-------------+-------------+-----------+
| 4        |     1.8439  |     1.34307 | 0.987478  |
+----------+-------------+-------------+-----------+
| 5        |     1.76848 |     1.28474 | 0.988482  |
+----------+-------------+-------------+-----------+
| Average  |     1.82139 |     1.32912 | 0.987863  |
+----------+-------------+-------------+-----------+
```

Figure 2 Best Model after optimization

4. Challenges and Limitations

Despite the promising results and successful deployment of the predictive system, several technical, logistical, and contextual challenges emerged during the model development, evaluation, and deployment phases. These challenges introduced complexities that required thoughtful handling and underscore areas for future improvement.

One of the foremost challenges encountered in this project was related to the intrinsic accuracy and operational stability of the locally constructed analog gas sensors. Although these sensors offer a low-cost and accessible solution ideal for deployment in resource-constrained environments, they inherently suffer from a range of technical limitations. These include susceptibility to long-term signal drift, which gradually alters their baseline readings over time, and high sensitivity to environmental factors such as temperature, humidity, and the presence of multiple interfering gases. Such cross-sensitivity often causes the sensor output to reflect not just the target gas (CO_2), but also unrelated gases or atmospheric phenomena, thus introducing noise and ambiguity into the data.

In addition to sensor-specific issues, the environmental conditions in which the data was collected posed another set of challenges. Although the dataset provided broad coverage of urban and peri-urban Nigerian climates, most of the readings were concentrated within the tropical zones typical of sub-Saharan Africa. This climatic uniformity poses a potential risk to model generalizability, as performance may degrade when the system is deployed in significantly different climates such as high-altitude, arid, or heavily industrialized regions where environmental variables diverge markedly from the training set.

Another important limitation stems from the nature of the dataset itself. The publicly available test set provided on the Zindi platform lacked access to the ground-truth CO_2 concentration labels. Consequently, while the model could be trained on labeled data, final evaluation on the test set had to rely on Zindi’s closed scoring mechanism, limiting direct insights into real-world prediction accuracy. However, the model's strong external validation score on Zindi's leaderboard indicated that it successfully generalized beyond the training distribution, thus serving as a proxy for field-level performance.

Hardware constraints represented yet another key area of limitation. The three Nigerian-manufactured sensor devices (Alpha, Beta, and Charlie) used for data collection were built using cost-effective microcontroller platforms that lack the computational power necessary to run inference models locally. This constraint necessitated the use of a cloud-based deployment architecture, wherein sensor data is transmitted to a centralized server for prediction. While this approach offers scalability and performance advantages, it introduces dependency on stable internet connectivity, which may be unreliable or unavailable in rural or underserved areas. This limitation highlights the trade-off between system affordability and infrastructure requirements.

Furthermore, while preprocessing and feature engineering mitigated some of the sensor inconsistencies, differences in hardware calibration, gas sensitivity thresholds, and analog signal resolution across devices introduced subtle discrepancies in the input data. These disparities necessitated the development of a device-aware validation strategy to ensure that model performance remained stable regardless of the device used for data acquisition.

Finally, although the deployed web application includes functionality for retraining the model with newly acquired data, this retraining still requires manual intervention and is not fully automated. This limits the model’s ability to adapt in real time to evolving environmental conditions, emerging emission sources, or progressive sensor wear-and-tear over prolonged deployments.

In summary, while the project has established the feasibility and potential of machine learning-enhanced CO_2 prediction using low-cost, locally built sensors, it also uncovered meaningful limitations that must be addressed to enable large-scale, long-term, and autonomous deployment. Tackling these challenges in future iterations will be essential for building a more adaptive, resilient, and universally deployable system for environmental monitoring.

5. Recommendations and Next Steps

To build upon the foundational success and practical utility demonstrated by this project, several forward-looking recommendations and concrete next steps have been identified. These proposals aim to expand the solution's applicability, reinforce its robustness, enhance its accessibility, and ultimately maximize its environmental and societal impact.

First and foremost, a critical recommendation involves significantly broadening the geographical and environmental scope of sensor deployments. Currently, most data feeding into the model originates from tropical climates typical of sub-Saharan Nigeria. While this has proven sufficient for the pilot stage, future iterations must include sensor installations in more varied ecosystems and climatic zones. For example, deploying the sensors in regions with arid, semi-arid, high-altitude, industrial, and coastal characteristics would enable the collection of more diverse and representative environmental data. This diversity would directly contribute to improving the generalizability and adaptability of the machine learning models, reducing the likelihood of bias or overfitting to a narrow range of environmental inputs.

Second, it is highly recommended to integrate a fully automated model retraining and deployment pipeline into the web application framework. While the current system supports manual retraining, automating this process would ensure that the model continuously evolves as new sensor data becomes available. Incorporating mechanisms such as scheduled retraining jobs, data drift detection, and automatic evaluation of model updates could ensure that the predictive accuracy remains high over time, even as sensor performance degrades or environmental conditions change. This continuous learning capability is vital for maintaining relevance and precision in real-world, long-term deployments.

Third, while cloud-based deployment has proven effective in supporting the computational demands of model inference and data processing, the reliance on internet connectivity limits the platform's usability in remote or underserved locations. To address this, the next phase should explore the feasibility of deploying lightweight, compressed versions of the model directly on edge devices. Using techniques such as model pruning, quantization, and hardware-aware neural architecture search, it is possible to adapt the models for on-device inference on low-power microcontrollers. This advancement would open the door to fully offline CO_2 prediction systems, particularly useful in agricultural, rural, and disaster-prone areas with limited connectivity.

Fourth, it is essential to engage local communities, educators, public health institutions, and governmental stakeholders in awareness-building and capacity-building efforts. Outreach programs such as environmental monitoring workshops, school-based pilot deployments, and participatory citizen science campaigns can democratize access to air quality data and empower communities to interpret and act on it. Creating public dashboards or integrating the tool into educational curricula can help nurture an informed and climate-conscious population that values data-driven decision-making.

Fifth, establishing partnerships with environmental regulatory bodies, climate-focused NGOs, and municipal planning agencies is key for scaling the real-world impact of this solution. By providing verifiable, high-resolution, and locally sourced CO_2 data, the system can become an indispensable tool for informing urban policy, shaping emission reduction strategies, guiding industrial zoning, and influencing health risk assessments. Aligning the platform with national climate action frameworks and international commitments such as the Paris Agreement or SDG 13 (Climate Action) will further amplify its policy relevance.

Finally, a long-term recommendation is to foster a collaborative ecosystem around this platform one that includes hardware innovators, software developers, data scientists, environmental researchers, and community stakeholders. Open-sourcing the sensor firmware, pre-processing code, and model architecture (with appropriate licensing) can invite broader contributions and accelerate innovation. Supporting this with transparent data sharing protocols and version-controlled APIs will ensure sustained growth and community trust.

In conclusion, while the current deployment serves as a powerful proof of concept, transforming it into a long-lasting, scalable, and impactful environmental intelligence platform requires strategic expansion, technological evolution, and strong community and policy integration. Through these recommended steps, the solution can play a pivotal role in strengthening Africa’s capacity for environmental monitoring and climate resilience.