Practical Machine Learning Final Assignment

Executive Summary

Our project targets to predict the annual direct point greenhouse gas (GHG) emissions from industrial facilities before they are officially reported. This prediction model will serve three primary business purposes: understanding industry trends, informing policymakers, and validating reported emissions. We will use EPA direct point GHG emissions data from industrial facilities from 2011 to 2023 to build our machine learning model. Success will be measured by the accuracy of our emissions predictions, evaluated through cross-validation of test-train datasets.

Group Members

- David Van Dyke
- Ethan Norton
- Migus Wong
- Michael Yoo

BUSINESS OBJECTIVE

The goal is to predict the annual direct point GHG emissions from industrial facilities before they are officially reported. This prediction model will serve the following business purposes:

- Understand Industry Trends: Identify future changes in emissions levels to guide potential investment strategies.
- Inform Policy Makers: Deliver accurate, timely predictions that enable regulators to shape future environmental policies.
- Resource Optimization and Cost Reduction: Flag facilities that deviate from industry norms so that investments and regulatory focus can be directed to those areas, reducing unnecessary costs
- Validate and Forecast Emissions: Serve as an independent benchmark to assess the accuracy and reasonableness of reported data, thereby informing risk assessments and ESG evaluations.

PROBLEM STATEMENT

Industrial facilities must report their GHG emissions if they meet or exceed a threshold (typically ≥25,000 metric tons CO_2 per year). However, relying solely on end-of-year reported data can delay critical interventions. More importantly, there is a lack of regulation overall within the energy sector. Regulation is primarily focused on the emission of harmful emissions, but not all of them. Our task is to build predictive models that estimate facility emissions in advance, using historical data and engineered features. This will empower stakeholders—including investors and policymakers—to act proactively rather than reactively.

RESEARCH OBJECTIVES

Primarily, our goal is to develop a machine-learning pipeline that automates data ingestion, cleaning, feature engineering, and model training, as well as analyze trends and industry differences to detect anomalies and forecast future emissions. Ultimately, we want to provide actionable insights through comprehensive visualizations and analysis for policymakers and investors.

Success will be measured by predictive accuracy, validation, and compliance detection. R^2 scores will assess how well the model captures the variability of emissions and RMSE to

quantify prediction error. Validation will be evaluated by K-fold cross-validation to ensure consistency in model performance using different subsets of data. Lastly, compliance detection will be measured through precision and recall. High recall is key in catching potential non-compliant facilities, while high precision mitigates false alarms.

EDA / DATA PREPARATION / FEATURE ENGINEERING

We will leverage annual greenhouse gas (GHG) emissions data from industrial facilities—collected from 2011 through 2023—to build our machine-learning models. This dataset originates from the EPA's Greenhouse Gas Reporting Program (GHGRP), which requires facilities and suppliers to report their emissions if they exceed defined thresholds—typically 25,000 metric tons of CO_2 per year for most industries. Since 2010, the EPA has published annual spreadsheets summarizing these reported emissions, using a consistent Facility ID as the primary key to track each facility's data over time. We evaluated the multi-year summary of EPA GHG emissions from direct point emitters (measured in metric tons) for our analysis. Our data sources include:

- Multi-Year Data Summary: High-level information for facilities over multiple years.
- Yearly Spreadsheets: Detailed yearly information, including reported emissions by greenhouse gas and process.

EDA/FEATURE ENGINEERING

Below is a table summary of the team's EDA and data preparation. Please reference our Python notebook if you'd like more details. However, we also highlight some key findings that caught our attention below.

EDA	TASK	VISUALS
Missing Data Analysis	•Checked Null values and dropped columns w/ >70% •Created table.summary logs for missing counts	Python print statement

Yearly Emissions Stats	• Calculated facilities count per/yr • Mean + total emissions • Observed trends	Table/Summary Logs
State Level Analysis	• Grouped data by state to compare total & mean emissions • Compared facility counts and emissions	Line plots
Outlier Detection	• Identified top 100 • Filtered data for deeper review	Box/Violin plot
Top 100 Emitters	• Ranked Facilities by total emissions • Summarized by state and industry sector	Bar Charts Heatmaps
Monitoring Status Exploration	• Compared emissions for continuous monitoring • Conducted t-tests for emission differences	Box and Violin Plots t-test summaries
Industry/Sector Analysis	• Examined Industry Type (sector) for top contributors • Created dummy variables for correlation check	Bar Charts Correlation Heatmaps
Modeling (Random Forest & Decision Tree)	• Encoded categorical columns • Trained ML models to predict emissions • Checked feature importance	Feature Importance Bar Plot Actual vs Predicted Scatter

MODELS IMPLEMENTED AND BUSINESS RELEVANCE

Multiple machine learning models were developed and evaluated to address the business problem. Each model aligns with the unique business cases.

- Linear regression - establish a clear benchmark for emissions forecasting, we develop a simple, interpretable model trained using cross-validation. This ensures reliability while keeping it accessible for policymakers and stakeholders to make informed decisions.

- Decision Tree Regressor - captures non-linear relationships in emissions data, where we used grid-search to test parameters like “max_depth” or min_samples_split to get the right level of complexity. This model gives detailed insights into what’s driving emissions, mainly to help regulators figure out compliance strategies.
- Random Forest Regressor - aims to improve predictive accuracy through ensemble learning. We achieved this by optimizing hyperparameters using RandomizedSearchCV with 3-fold cross-validation and testing generalizability through out-of-bag (OOB) validation. This approach results in more reliable and generalizable predictions, benefiting investment analysis and policy-making decisions.
- Gradient Boosted Trees - we utilized sequential boosting to enhance accuracy beyond Random Forest. A grid search optimizes key hyperparameters (n_estimators, learning_rate, tree depth) with 5-fold cross-validation to prevent overfitting. This approach delivers precise emissions forecasts, aiding industry regulations and risk assessment.
- Neural Network - capture complex, non-linear relationships in emissions data, we applied PCA for dimensionality reduction, comparing models with 101 and 401 features. The Adam optimizer optimized the network with early stopping to prevent overtraining. This improves accuracy for ESG compliance and long-term forecasting, though interpretability remains challenging when communicating results to stakeholders.
- Clustering for Industry Segmentation - group facilities with similar emissions patterns, we explored K-Means (k=4), DBSCAN, and Agglomerative Clustering, using silhouette scores to evaluate clustering quality. This approach helps regulators and investors pinpoint high-emission sectors, enabling more focused policies and strategic investment decisions.
- Anomaly Detection for Compliance Monitoring - detect potential non-compliance in emissions reporting, we applied outlier detection techniques, including Isolation Forest, One-Class SVM, and autoencoder-based reconstruction error analysis. This improves the identification of under-reporting or abnormal deviations, strengthening environmental compliance and risk management efforts.

FINDINGS AND CONCLUSIONS

RANDOM FOREST WITH PCA

The project built a random forest model (a bunch of decision trees working together) to analyze emissions data. Since there were hundreds of categorical features, the team used PCA (Principal Component Analysis) to shrink them to about 100 key components before training 71 trees (each up to 30 levels deep, requiring at least 4 samples per split). This kept ~95% of the variance while making the model more manageable. The model first ran R^2 of ~0.27 and a 5-fold cross-validation R^2 of ~0.33. It reduced error (RMSE) to ~1.1 million emissions units, improving over the baseline 1.31 million (the standard deviation of emissions). Predictions tended to follow the expected trend but had some underestimation for extreme emitters. The most critical factors influencing emissions were industry type (e.g., NAICS 221112, likely fossil fuel power plants) and state location, accounting for 64% of the model's decision-making importance. Other influential variables included whether a facility uses continuous emissions monitoring and the Year of operation, showing that emissions change over time due to compliance efforts and industry trends. The model provides industry insights, highlighting which sectors and states contribute the most emissions. This helps stakeholders track trends and make data-driven decisions. For policy guidance, regulators can focus on high-emission sectors, such as fossil fuel power plants, and use the model's insights to set more effective environmental policies. From a resource allocation standpoint, agencies and businesses can target enforcement and mitigation efforts where they'll have the most significant impact, optimizing monitoring budgets.

Additionally, the model is helpful for validation and forecasting. It can flag facilities with unexpected emissions, helping detect reporting errors. It also estimates future trends based on past patterns, providing scenario-based forecasts rather than full time-series predictions.

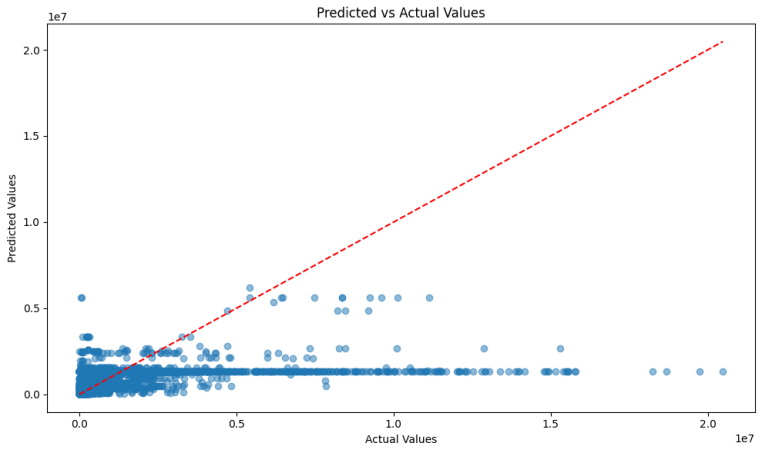
NEURAL NETWORK WITH PCA INPUTS

One deep learning approach was a fully-connected artificial neural network that processed PCA-compressed features. The network was a multi-layer perceptron with four hidden layers (64-128-64-32 neurons), using ReLU activations and dropout regularization to prevent

overfitting. PCA reduced the original 416 features to 101 principal components, simplifying the input space while preserving key patterns. Dense layers then learned complex nonlinear relationships in emissions, with early stopping applied to prevent over-training.

The model achieved a test R^2 of ~0.20 (explaining ~19.7% of variance) and a mean absolute error (MAE) of ~417k. Training stopped at epoch 74, with early stopping triggered at epoch 59 when validation loss plateaued. Despite capturing some patterns, its RMSE (~1.16 million) lagged behind the random forest, suggesting PCA may have removed practical industry-specific details. Predicted vs. actual emissions showed general alignment but significant scatter, particularly for large emitters, indicating underestimation of extreme cases. Training history revealed improving loss trends but an initial gap between training and validation loss, hinting at underfit capacity or information loss from PCA.

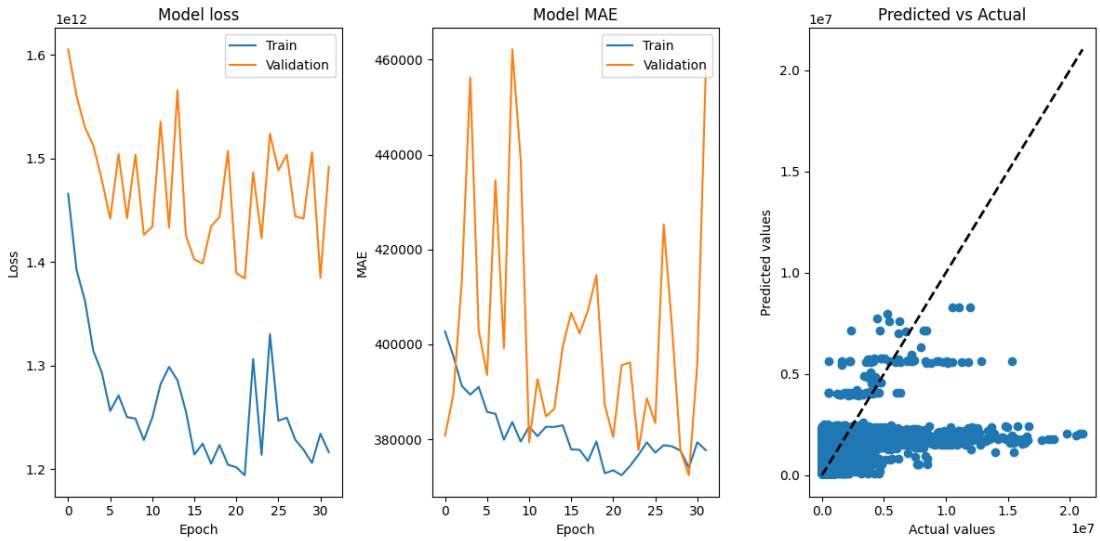
While this neural network contributed to understanding industry trends, its reliance on PCA made direct feature interpretation difficult, limiting its transparency for policymakers. However, it still provided insights—consistent underestimation of certain facilities suggested missing factors, such as production volume or technology use. As a proof-of-concept, its moderate accuracy suggests deep learning has potential for emissions modeling but requires additional refinement before guiding resource allocation. For validating and forecasting emissions, the model could flag reporting discrepancies and generate emissions projections based on hypothetical scenarios, though PCA's abstraction constrained its effectiveness. Despite underperforming compared to other models, this approach highlighted areas for future improvements, such as incorporating more granular industry data to enhance deep learning's applicability.




```
Epoch 78/200
2065/2065 ━━━━━━━━━━━━━━━━━━━━ 6s 3ms/step - loss: 1332086439936.0000 - mae: 409385.4062 - val_loss: 1530066370560.0000 - val_mae: 430655.9375
Epoch 79/200
2065/2065 ━━━━━━━━━━━━━━━━━━━━ 5s 3ms/step - loss: 1448503279616.0000 - mae: 423273.8438 - val_loss: 1533300441088.0000 - val_mae: 425389.4688
Epoch 79: early stopping
Restoring model weights from the end of the best epoch: 64.
443/443 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - loss: 1365452128256.0000 - mae: 414098.3750
Test loss (MSE): 1340960145408.00
Test MAE: 414132.41
443/443 ━━━━━━━━━━━━━━━━━━━━ 1s 1ms/step
Test R-squared: 0.2003
```

NEURAL NETWORK WITHOUT PCA

The team built a deep feed-forward neural network that used the complete feature set instead of PCA-reduced components. This meant the model considered all categorical indicators (industry sectors, yes/no flags, state, year, etc.), increasing input size but preserving detailed information. L2 regularization, dropout (20–30%), and batch normalization were applied to manage overfitting, a learning rate scheduler and early stopping (patience ~10 epochs). The neural network slightly outperformed the PCA-based version, achieving an R^2 of ~0.25–0.26 across all data splits (train ~0.25, val ~0.23, test ~0.256), showing consistent generalization. Test RMSE was ~1.12 million (vs. 1.30 million baseline), and MAE ~388k, meaning it captured more variance than the PCA-network. However, its accuracy was similar to the random forest (slightly lower R^2) and below the gradient boosting model. The predicted vs. actual emissions plot showed better clustering along the diagonal than the PCA-network but still had some scatter for high emitters. Unlike tree models, feature importance wasn't directly available, though future analysis using permutation importance or SHAP values could provide insights. Comparing performance with the PCA version suggested the neural net captured specific industry codes and state effects that PCA had averaged out.



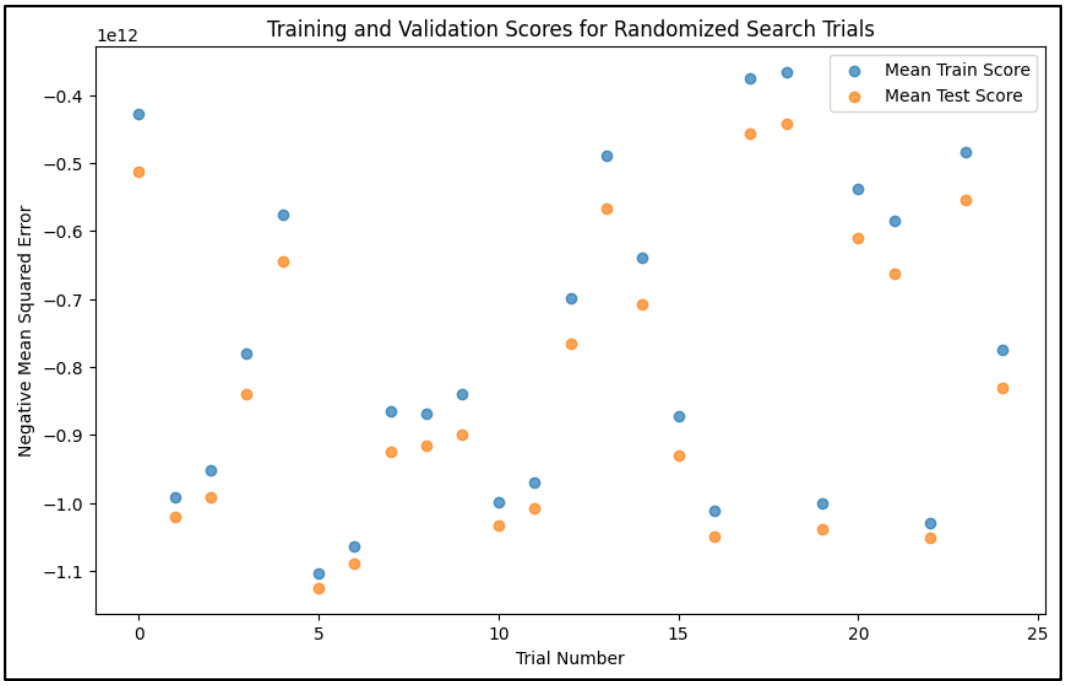
The model helps identify emission patterns across sub-industries and states for business and policy, offering more profound insights into nonlinear behaviors. Since it directly uses original features, it can also simulate policy scenarios—for example, predicting how emissions might drop if facilities adopt carbon capture technology. While less interpretable than tree models, its ability to capture complex feature interactions makes it helpful in detecting risk factors for emissions spikes, which can guide strategic audits or interventions. Lastly, it serves as a tool for validating unexpected emissions and forecasting future trends, supporting long-term planning. Finally, though less explainable than tree-based models, the neural network's improved performance highlights the importance of retaining granular industry data. Businesses should invest in comprehensive data collection and consider interpretability tools to make deep learning models more actionable.

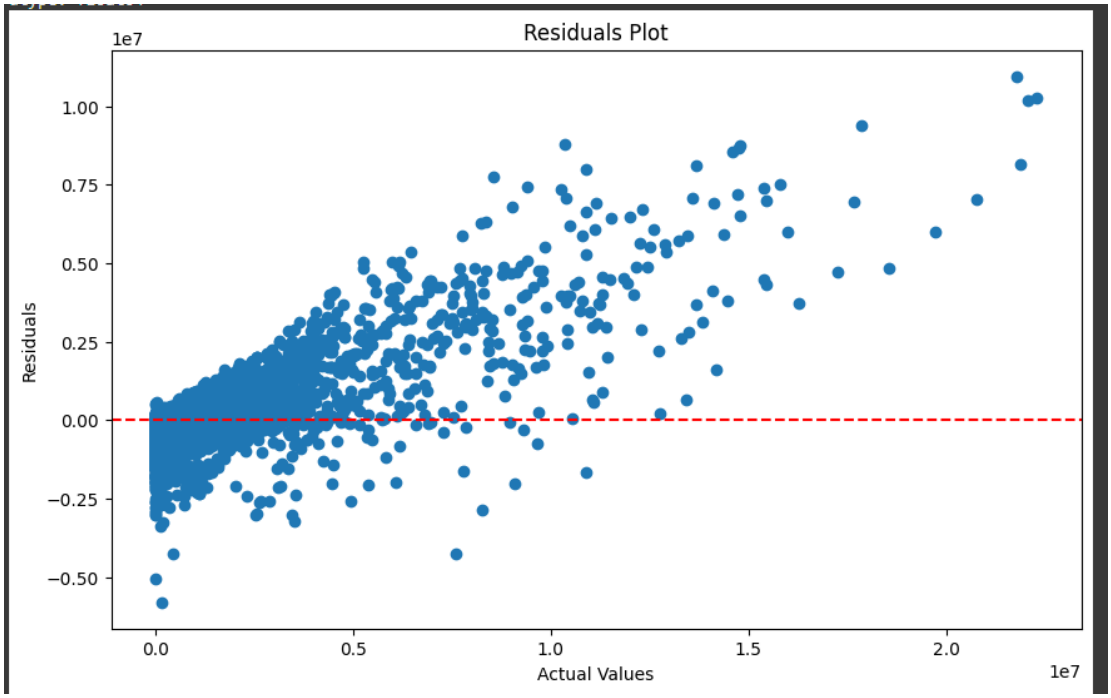
GRADIENT BOOSTED TREE MODEL

The gradient boosted trees model was the most accurate and reliable in this project, achieving R^2 ~0.75 and RMSE ~650,375, significantly outperforming both neural networks and the random forest. It used 239 trees, a max depth of 5, and a learning rate of ~0.16, tuned through grid and randomized search. Unlike PCA-based models, it efficiently handled categorical features without dimensionality reduction, leveraging subsampling (0.84) and regularization to prevent overfitting. A 5-fold cross-validation ensured robust performance before final testing. Predictions closely matched actual emissions, with key drivers including NAICS industry codes,

state, and continuous emissions monitoring, reinforcing findings from other models but with superior predictive accuracy.

This model provided valuable insights for industry trend analysis, policy forecasting, and resource optimization. Its high accuracy meant analysts could trust its predictions to reveal nonlinear emissions patterns, such as thresholds where emissions spike as production increases. Policymakers could use the model to simulate the effects of regulatory changes, such as predicting how stricter emissions limits or new technologies might impact future emissions. Unlike other models, its precision allowed for scenario-based decision-making, offering quantitative estimates for different policy actions.





Beyond policymaking, the model was a resourceful tool for resource allocation. It pinpointed which factors most influence emissions, helping businesses and regulatory agencies prioritize high-impact mitigation efforts while avoiding unnecessary spending on low-impact interventions. Additionally, it was instrumental in validating reported emissions and forecasting future trends. Significant deviations between reported and predicted emissions could indicate data errors or unusual activity, prompting further investigation. Its strong forecasting ability made it ideal for long-term climate planning, generating state- and national-level projections to ensure emission reduction targets stay on track

LESSONS LEARNED AND RECOMMENDATIONS

Understanding the big picture, greenhouse gas emissions have declined significantly from 3.2 billion metric tons CO_2e in 2010 to 2.4 billion in 2023, with per-facility emissions dropping from ~507,600 to ~368,300 metric tons. This reflects efficiency gains, cleaner technology, and plant retirements (U.S. Environmental Protection Agency n.d.). Power generation remains the top emissions source, but levels have declined due to the shift from coal to gas and renewables. In contrast, chemical manufacturing emissions have increased, suggesting industry growth or slower mitigation efforts, while waste management facilities contribute less per site. Emissions

are concentrated in industrial-heavy states like Texas, Louisiana, and Indiana, where power plants, refineries, and factories dominate. Many high-emission states have seen declines since 2014, likely from plant closures or upgrades, though some regions with growing industries may not follow this trend. Power plants and large industrial sites are the most significant contributors, with NAICS 221112 (fossil fuel power generation) having the strongest correlation (~0.33) to emissions. Facilities engaged in both power generation and industrial production also show exceptionally high emissions. While emissions are trending downward, sector and regional differences highlight the need to target high-emitting industries and locations for maximum impact.

For informing policymakers, the data highlights power generation, petroleum refining, and large industrial manufacturing (chemicals, metals) as the highest-emitting sectors, making them prime targets for regulation and policy support. These industries contribute the most to total emissions and operate the most extensive facilities—some coal-fired power plants alone emit 10–16 million metric tons of CO_2e annually. Because emissions are concentrated in a few significant sources, targeted regulations, incentives for cleaner energy, and stricter emissions standards in these sectors could drive substantial reductions with relatively few interventions.

Examining sector and regional trends further reveals where emissions are declining, stagnant, or rising, helping policymakers focus efforts where reductions are most achievable. The steep decline in power sector emissions from 2010 to 2023 suggests that policies—such as cleaner electricity incentives and coal-to-gas transitions—have been effective. Expanding these efforts, including renewable energy deployment and efficiency programs, could sustain the downward trend. Meanwhile, industries with minimal declines or slight increases, like chemical manufacturing, represent opportunities for new interventions. Identifying what's driving these trends—outdated technology or weaker regulations—can guide sector-specific strategies like stricter emissions standards or innovation incentives. Similarly, regions with flat or rising emissions may benefit from state-level programs to modernize facilities and cut emissions.

The emissions data also indicates past policy effectiveness, with national emissions peaking around 2011–2014 before declining. This suggests that regulations implemented in the 2010s played a key role in reducing emissions, particularly in the power sector. Policies promoting cleaner electricity and industrial efficiency have demonstrated success, reinforcing the

importance of continuing and expanding well-designed regulations. At the same time, if specific sectors or regions do not respond as expected, adjustments may be needed to strengthen the policy impact.

Geographic patterns show that a handful of states—such as Texas, Louisiana, and Indiana—account for a disproportionate share of emissions, making them priority areas for regional policy targeting. Policymakers could tailor state-specific strategies to address dominant sources—for example, petrochemical emissions and methane leakage in Texas and Louisiana, versus coal plant retirements in other states. Conversely, states that successfully reduced emissions (such as those transitioning away from coal) could be models for best practices. Aligning policy interventions with the data ensures efforts are directed where they will have the most significant impact—focusing on high-emitting facilities and regions with emerging challenges.

For resource optimization and cost reduction, the data confirms that a small number of facilities contribute the majority of emissions, making it essential to prioritize high-emitting sites for regulation and mitigation. Large coal power plants like James H. Miller Jr. (AL) and Gen. J.M. Gavin (OH) each emitted over 13–16 million metric tons of CO_2e in a single year, with similar patterns observed in Texas, Indiana, Pennsylvania, and Illinois. By monitoring these top-tier emitters, agencies can address a large share of emissions with fewer interventions, making enforcement efforts far more cost-effective. Allocating more inspectors, frequent audits, or better monitoring equipment to these significant sources can yield higher returns on investment than spreading resources across numerous smaller facilities.

A critical gap in emissions monitoring was identified—only 11.6% of power plants use continuous emissions monitoring systems (CEMS), while 88.4% rely on periodic reporting or estimations. This lack of real-time tracking raises concerns about potential inaccuracies or underreporting. Many of the highest-emitting facilities still lack CEMS, highlighting an opportunity for regulators to prioritize these sites for compliance checks or mandate broader adoption of real-time monitoring. Investing in CEMS at key facilities would enhance transparency, improve enforcement efficiency, and help operators optimize emissions control, reducing costs over time.

Beyond monitoring, data science models were used to flag anomalies, helping identify potential non-compliance risks. If a facility's reported emissions are consistently lower than

expected, based on its characteristics and history, it may indicate underreporting or measurement issues. This allows regulators to focus limited enforcement resources on high-risk facilities instead of relying on random inspections. Analysis at the state level also revealed that states with large industrial footprints, like Texas and Louisiana, had more outliers, signaling areas where compliance enforcement should be strengthened.

The data suggests that capital investments in emissions reduction should be concentrated where it yield the most significant impact. Rather than distributing resources across many small sources, upgrading a handful of high-emitting facilities—such as retrofitting major power plants with carbon capture or advanced scrubbers—can significantly reduce overall emissions. Similarly, investing in better monitoring and data infrastructure enhances operational efficiency and regulatory oversight, preventing costly violations and enabling long-term cost savings. Since not all emissions sources are equal, focusing mitigation, monitoring, and investment efforts on the most significant emitters is the most effective way to achieve substantial and cost-efficient emissions reductions.

For validating and forecasting emissions, the machine learning models we developed to predict facility emissions, including decision trees, random forests, and neural networks with PCA, showed modest performance, reflecting the complexity of emissions forecasting. The random forest model achieved an R^2 of only 0.11–0.13, meaning it explained just 11–13% of emissions variance. This suggests that while models can identify broad patterns (e.g., distinguishing a power plant from a small factory), emissions fluctuations depend on operational factors not captured in the dataset, such as plant utilization, maintenance schedules, and technology differences. As a result, these models are better suited for trend analysis and anomaly detection rather than precise forecasting.

The analysis uncovered notable discrepancies by comparing predicted vs. reported emissions, highlighting potential underreporting or overestimation. For example, a university power plant in Florida was expected to emit ~26,100 tons, yet it reported only ~20,913 tons (a 20% lower figure), possibly indicating underreporting or exceptional efficiency. Conversely, a landfill in South Carolina reported ~86,955 tons in 2010, while the model predicted only ~56,930 tons, suggesting either a unique event or overestimation. These anomalies serve as flags for further investigation, whether to trigger audits, verify facility reporting methods, or improve model accuracy.

A significant insight was the low adoption of continuous emissions monitoring systems (CEMS)—only 11.6% of power plants used real-time monitoring, while 88.4% relied on periodic reporting or estimates. This gap suggests a higher risk of inaccuracies or intentional underreporting at facilities lacking CEMS. Prioritizing compliance checks at these sites or mandating wider CEMS adoption could improve data integrity and enforcement efficiency. The predictive analysis also helped detect underreporting risks, identifying facilities consistently reporting lower-than-expected emissions based on industry benchmarks. A facility reporting 30% less CO_2 than similar sites raises red flags for potential misreporting. While these anomalies do not confirm wrongdoing, they highlight cases that warrant closer inspection. State-level analysis showed that states with many large facilities, such as Texas and Louisiana, had more outliers, pointing to systemic reporting trends that may require policy intervention.

Despite these discrepancies, the historical emissions data was found to be reliable. The dataset contained 94,378 records from 2010–2023, covering 8,778 facilities, with no missing facility IDs. Integrity checks confirmed yearly trends aligned with known industry and policy shifts, suggesting the data accurately reflects real-world emissions changes. While anomalies require review, the dataset is a strong foundation for forecasting, regulatory decisions, and long-term policy planning. Expanding real-time monitoring and predictive validation can further improve data accuracy and compliance enforcement.

CITATIONS

U.S. Environmental Protection Agency. “GHGRP Emissions Trends.” EPA, n.d. https://www.epa.gov/ghgreporting/ghgrp-emissions-trends (accessed March 16, 2025).