



# Random Forest Classification of CO<sub>2</sub> Emissions

To complement my regression model, I applied classification to evaluate how different machine learning techniques interpret the same emissions signal.

With an R<sup>2</sup> of 0.9998, the regression model captured CO<sub>2</sub> levels with near-perfect precision. The classifier didn't replace it, it discretized the same signal to explore threshold logic and operational triggers, offering a complementary perspective for emissions monitoring.

# Overview

## Context

Understanding and managing CO<sub>2</sub> levels is critical for environmental monitoring and industrial emissions control

## Approach

Applied Random Forest models for both regression (continuous prediction) and classification (high/low detection), enabling multiple views of the same emissions signal

## Value

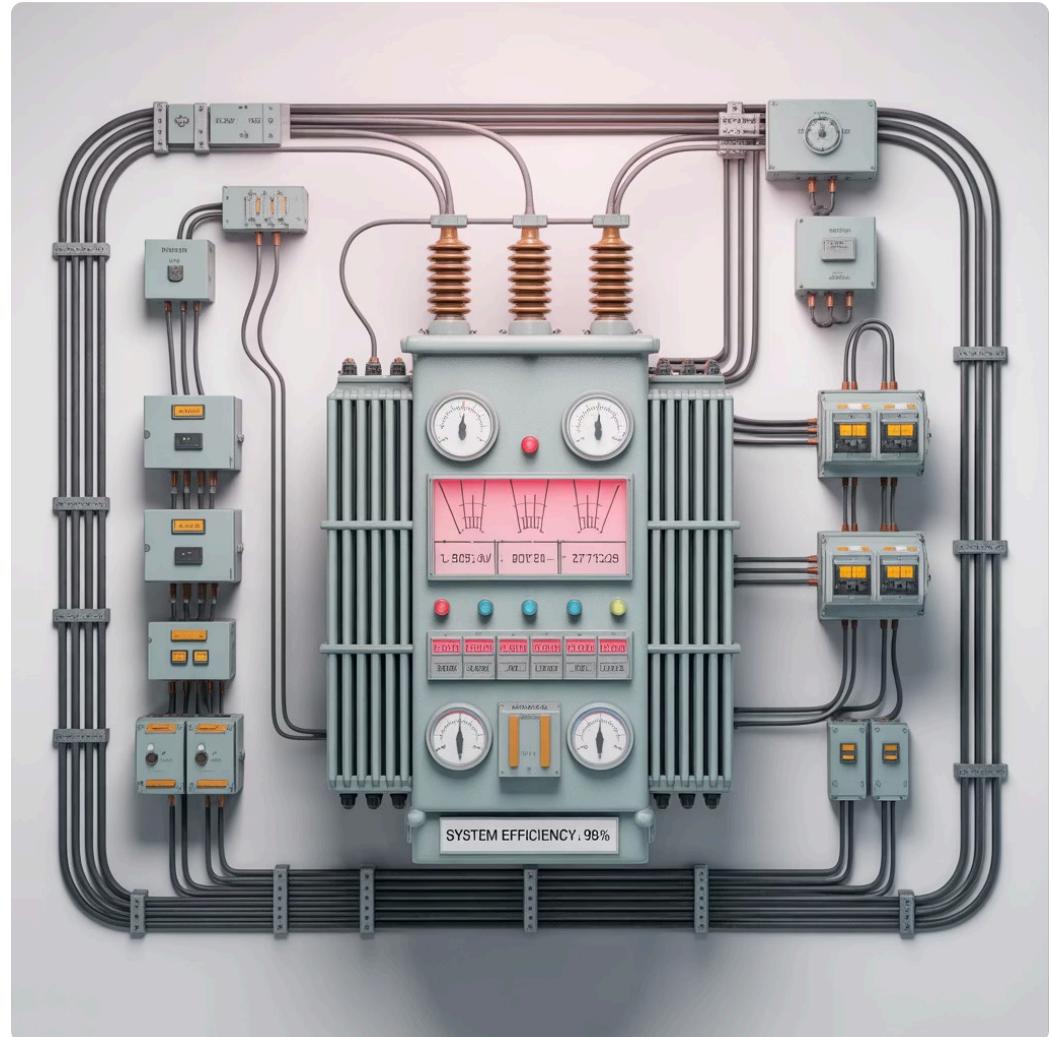
Delivered actionable insights through interpretable machine learning, identifying key drivers of CO<sub>2</sub> levels and supporting both forecasting and alert systems

# Key Features (Variables)

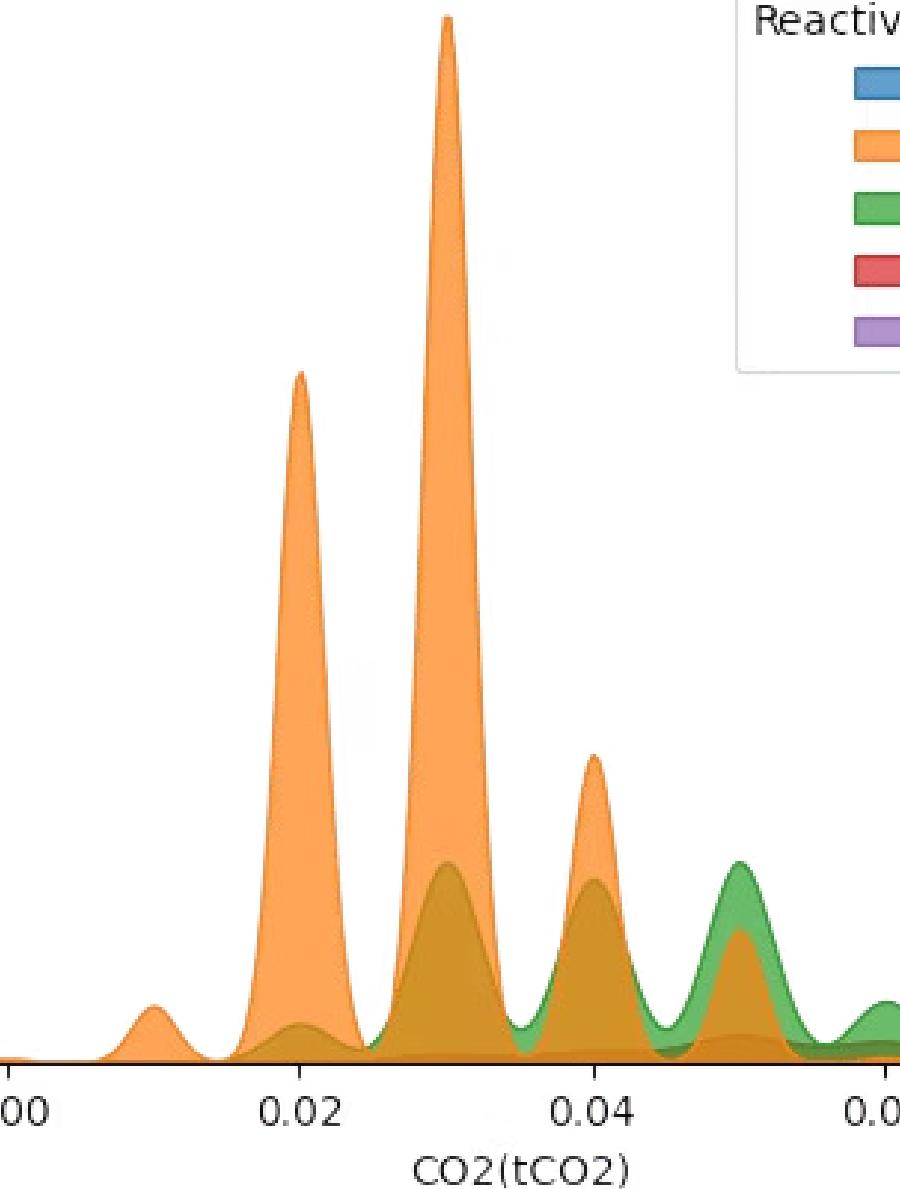
Both models leverage four critical features to predict and classify CO<sub>2</sub> concentrations:

- Lagging current reactive power
- Lagging current power factor
- Leading power
- Temperature

These features were selected based on their relevance to electrical system dynamics and energy efficiency. Details on feature selection and rationale are available in the main portfolio



## Density Plot Without 0-20 Bin



# Original Regression Model Performance

## RANDOM FOREST REGRESSOR

**0.998**

**R<sup>2</sup> Score**

Captures 99.8% of CO<sub>2</sub> variance, indicating exceptional predictive power

**0.0004**

**RMSE**

Extremely low root mean squared error demonstrates minimal prediction error

*Validation through residual and Q-Q plots (available in main portfolio) confirms well-behaved, unbiased errors, underscoring the model's reliability for continuous CO<sub>2</sub> prediction.*

# Classification Model Results

**99%**

## Accuracy

Correctly classifies 99% of 2,811 samples

Class 1 (High CO<sub>2</sub>) Metrics:

- Precision: 0.99
- Recall: 0.97
- F1 Score: 0.98

**0.99965**

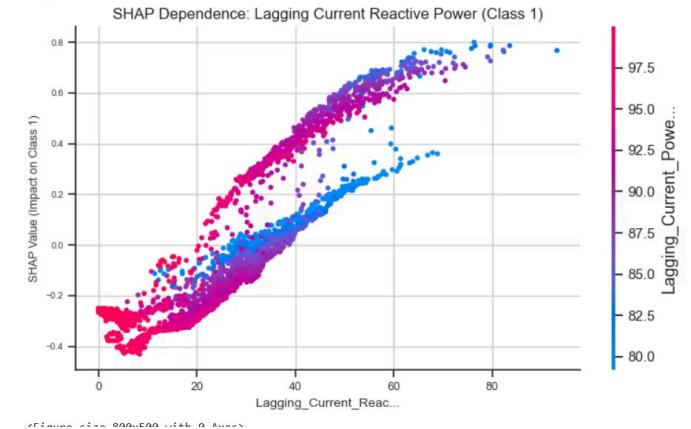
## ROC AUC

Near-perfect discrimination between classes

```
# Save SHAP dependence plot
plt.savefig(f"shap_dependence_{feature_name.lower().replace(' ', '_')}.png", dpi=400, bbox_inches='tight')
plt.show()
```

```
# Debug: Confirm feature names and data shapes
print("Feature names:", feature_names)
print("X_test shape:", X_test.shape)
print("shap_values shape:", shap_values.shape)
print("Mean |SHAP values| (Class 1):", np.abs(shap_values[:, :, 1]).mean(axis=0))
```

<Figure size 800x500 with 0 Axes>



# Confusion Matrix Breakdown

1

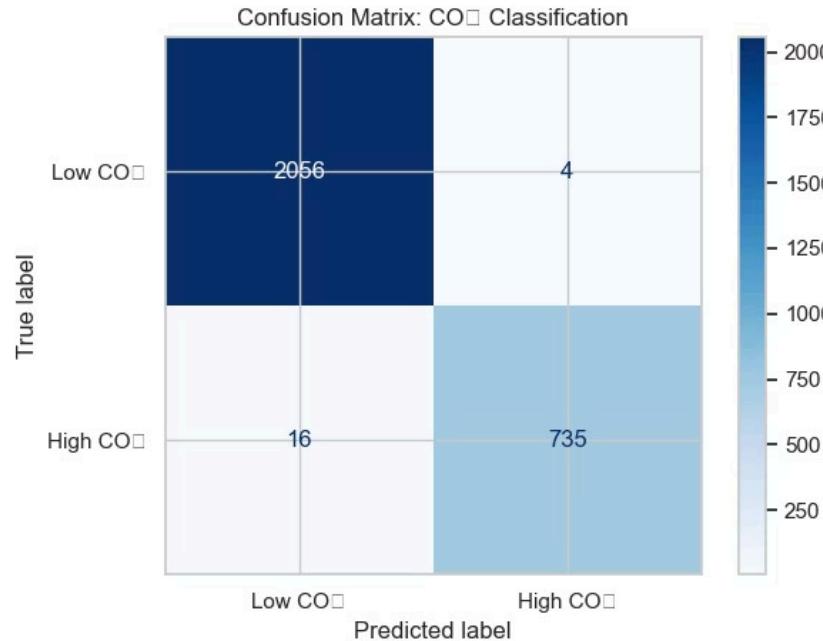
**True Positives: 735**

Correctly identified high CO<sub>2</sub> instances (Class 1)

2

**False Positives: 16**

Low CO<sub>2</sub> incorrectly classified as high



1

**True Negatives: 2,056**

Correctly identified low CO<sub>2</sub> instances (Class 0)

2

**False Negatives: 4**

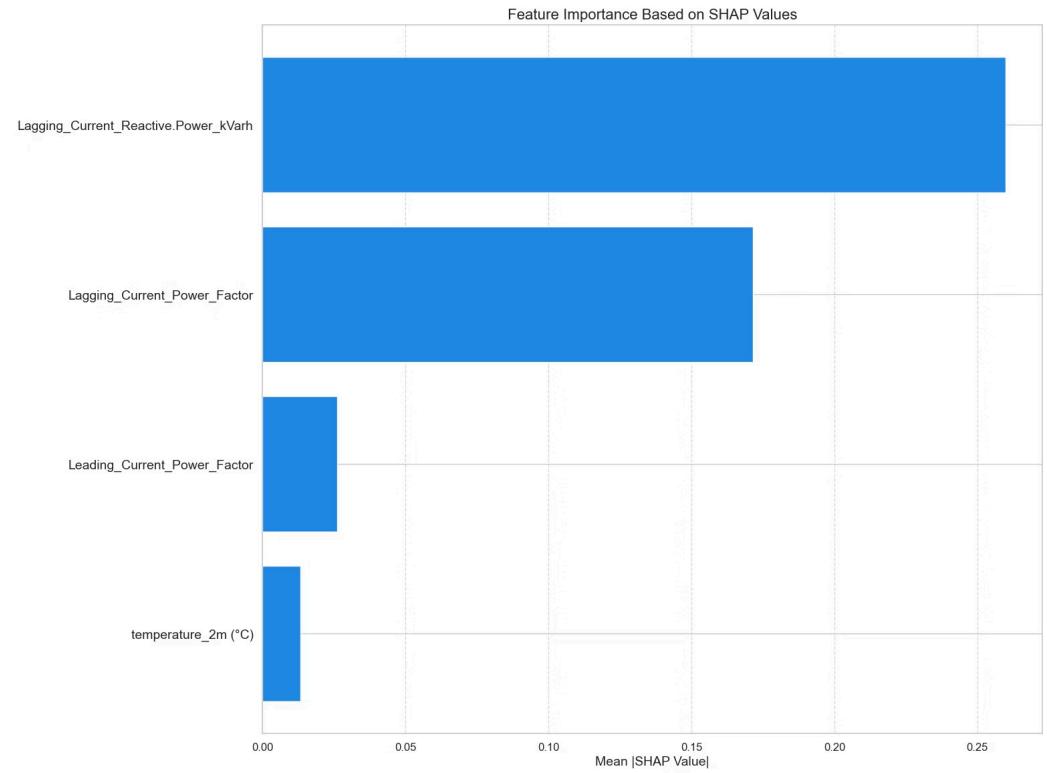
High CO<sub>2</sub> incorrectly classified as low

The model shows exceptional performance in identifying both high and low CO<sub>2</sub> conditions, with minimal misclassifications.

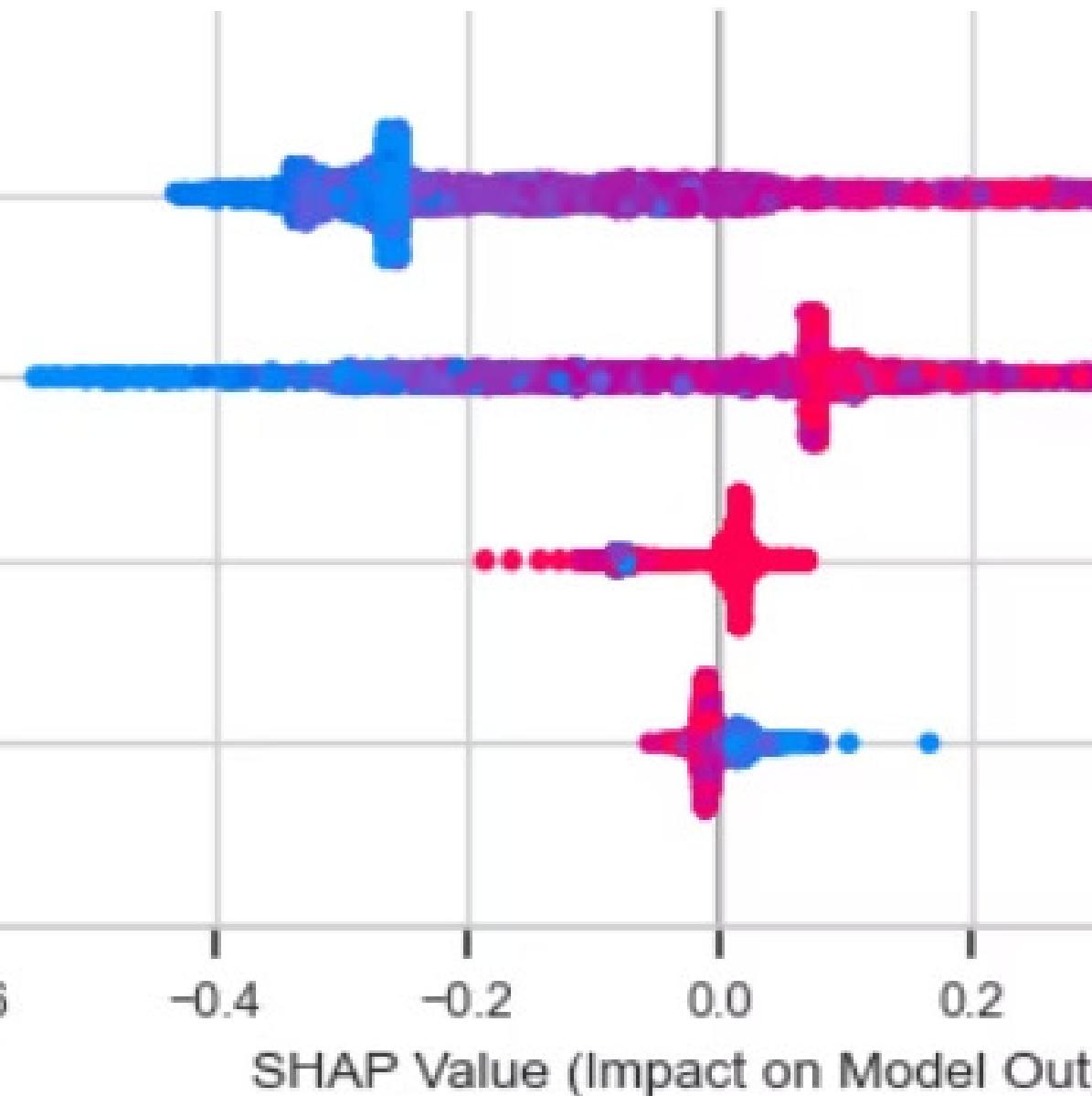
# SHAP Analysis: Feature Importance

## Beeswarm Plot Insights (Class 1):

- **Lagging current reactive power:** Dominant driver (SHAP range: -0.4 to +0.7)
- **Lagging current power factor:** Second strongest (SHAP: -0.6 to +0.45)
- **Leading power:** Moderate impact
- **Temperature:** Notable outlier boosting high CO<sub>2</sub> predictions



## SHAP Beeswarm Plot (Positive)

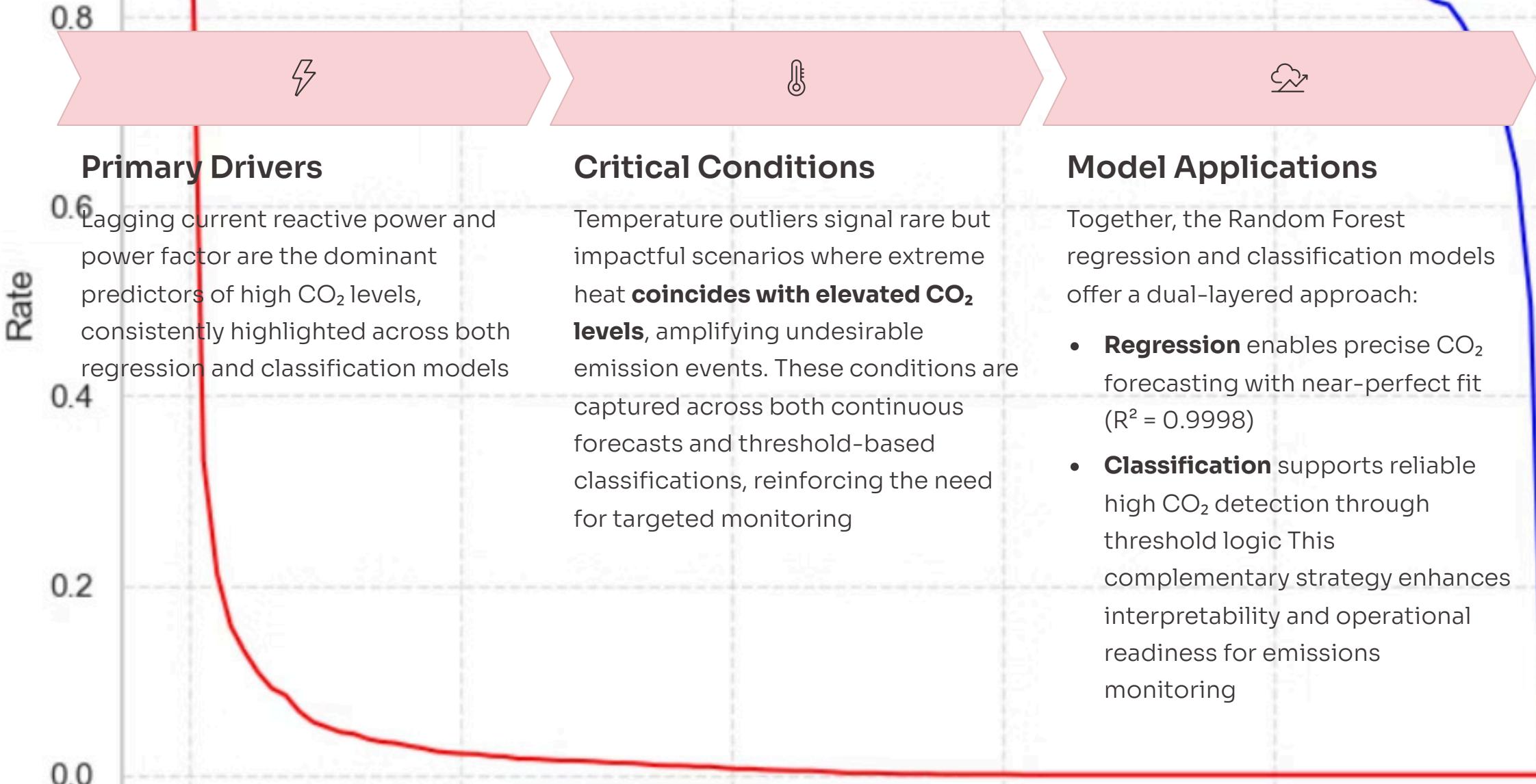


## SHAP Dependence Analysis

Dependence plots reveal critical relationships:

- High lagging current reactive power values sharply increase Class 1 probability
- Temperature plot highlights a high-value outlier, suggesting critical conditions where extreme heat amplifies CO<sub>2</sub> risks
- Clear thresholds identified where feature values significantly shift predictions

# Key Insights



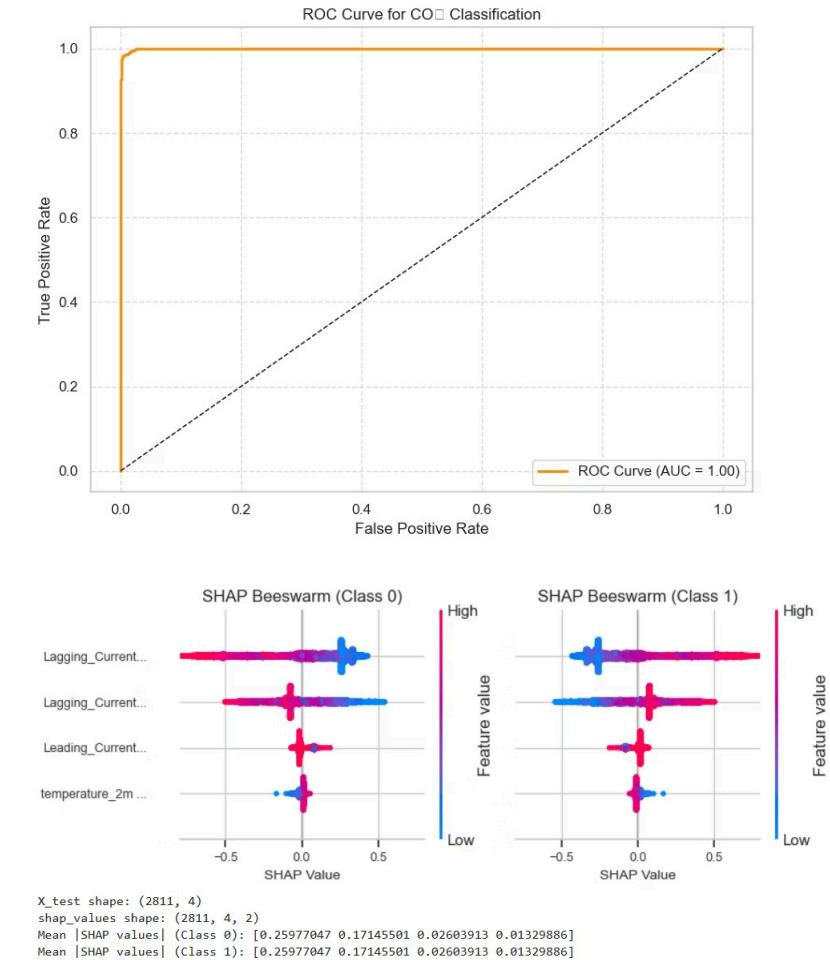
# Overview and Windup

## Conclusions

- Dual Random Forest models: regression and classification that successfully captured CO<sub>2</sub> dynamics
- Classification outperformed regression for high/low emissions detection, offering clearer operational thresholds
- Lagging current reactive power and power factor emerged as dominant predictors of elevated CO<sub>2</sub>
- SHAP analysis revealed interpretable feature impacts and critical temperature outliers influencing emissions

## Overall Findings & Learnings

- Machine learning can deliver both precision forecasting and actionable classification for emissions monitoring
- Feature engineering and model selection are pivotal, and classification unlocked clearer mitigation strategies
- SHAP visuals enhanced interpretability, making model logic transparent
- This project deepened my understanding of real-world modeling tradeoffs and diagnostic storytelling



*This work demonstrates proficiency in building, validating, and interpreting machine learning models, delivering actionable insights through advanced visualizations.*