Random Forest Classification of CO_2 Emissions

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

With an R^2 of 0.9998, the regression model captured CO_2 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_2 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

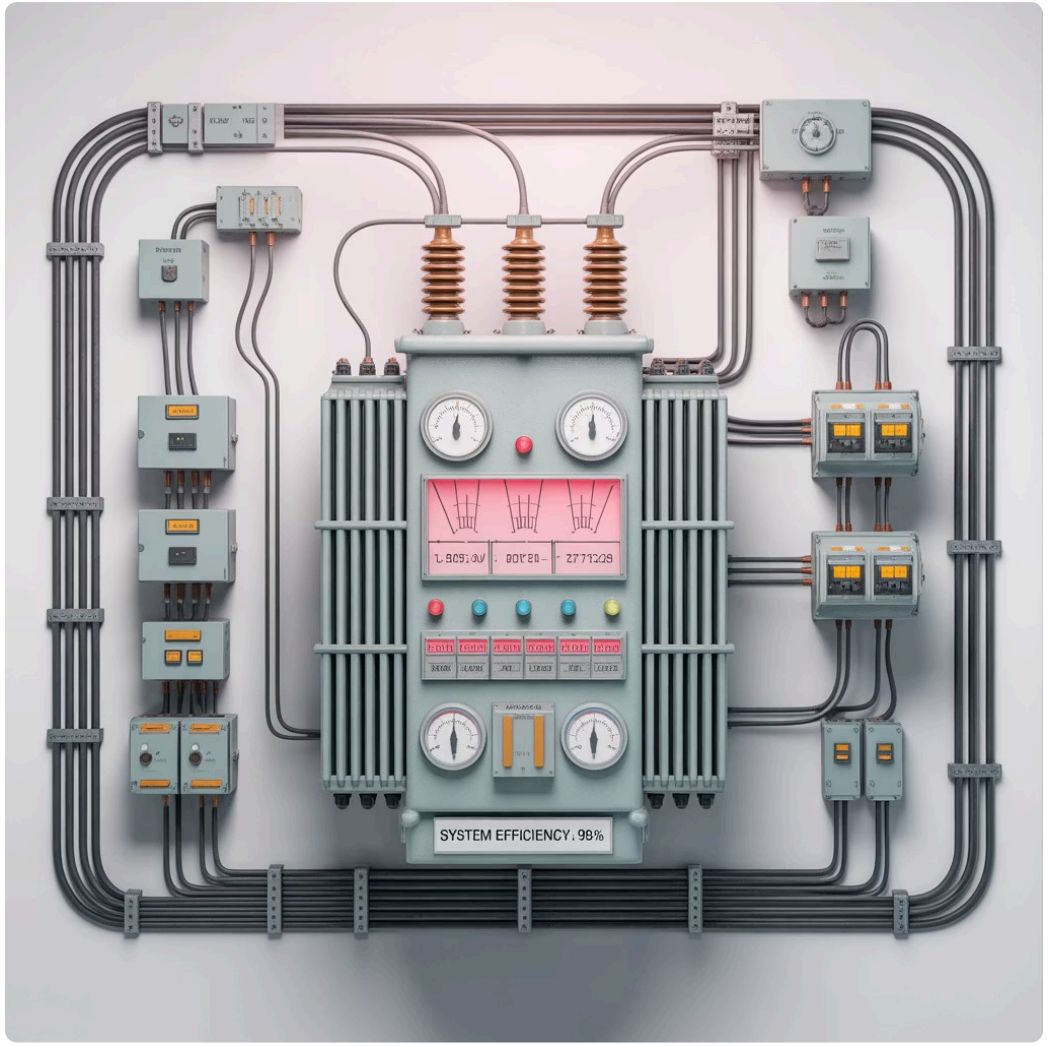
Delivered actionable insights through interpretable machine learning, identifying key drivers of CO_2 levels and supporting both forecasting and alert systems

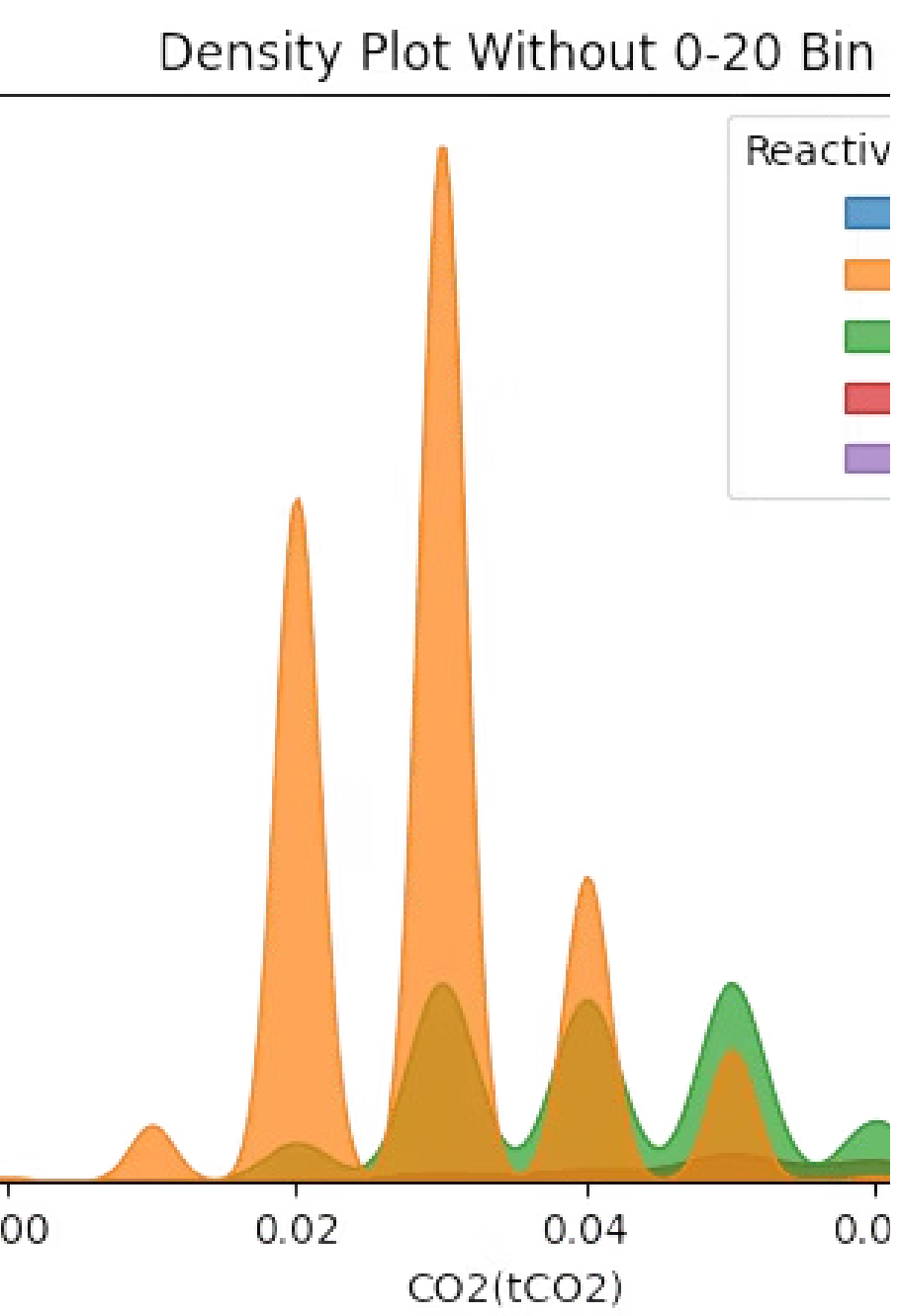
Key Features (Variables)

Both models leverage four critical features to predict and classify CO_2 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





Original Regression Model Performance

RANDOM FOREST REGRESSOR

0.998

R^2 Score

Captures 99.8% of CO_2 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_2 prediction.

Classification Model Results

99%

Accuracy

Correctly classifies 99% of 2,811 samples

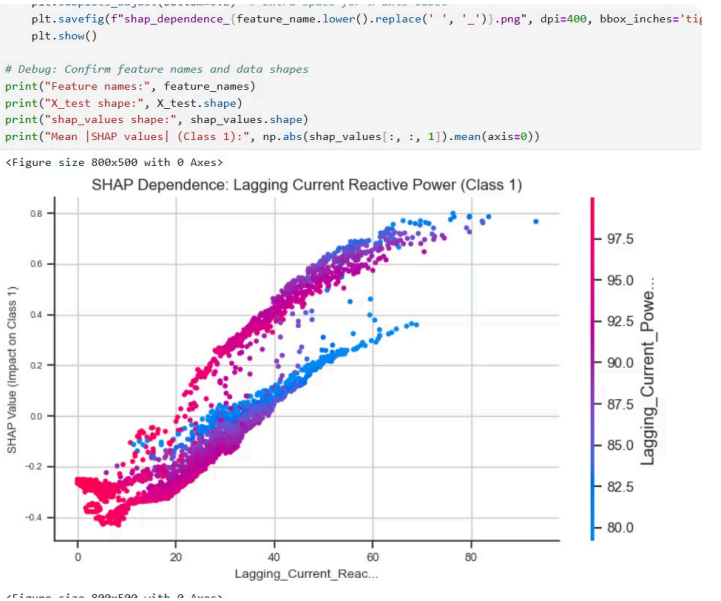
0.99965

ROC AUC

Near-perfect discrimination between classes

Class 1 (High CO_2) Metrics:

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



Confusion Matrix Breakdown

1

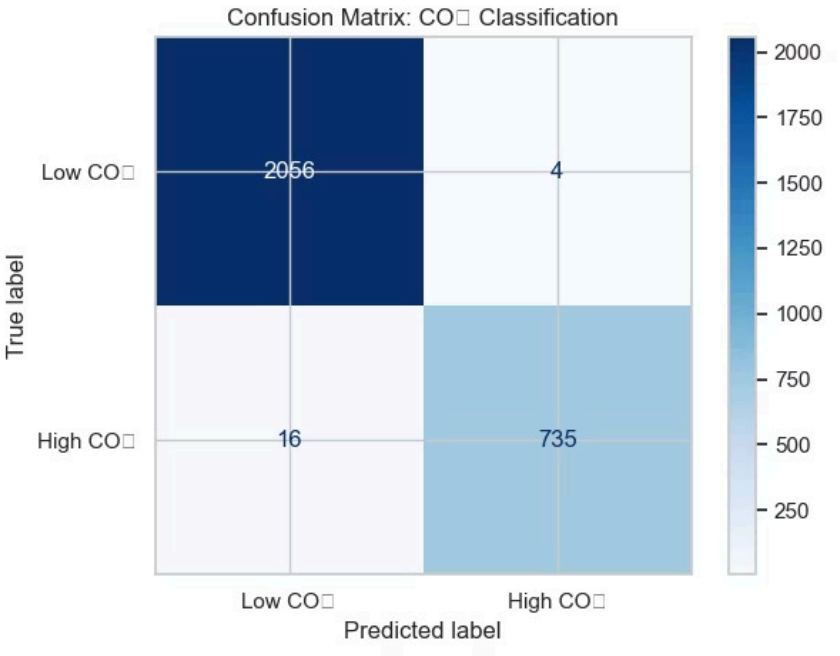
True Positives: 735

Correctly identified high CO_2 instances (Class 1)

2

False Positives: 16

Low CO_2 incorrectly classified as high



1

True Negatives: 2,056

Correctly identified low CO_2 instances (Class 0)

2

False Negatives: 4

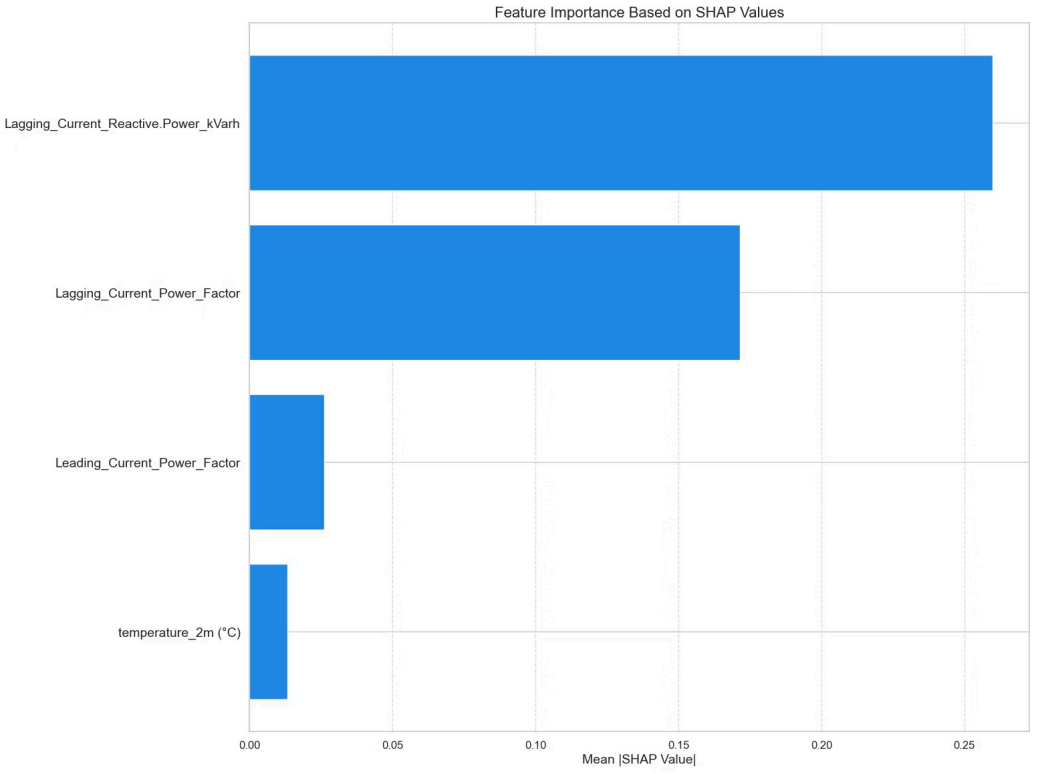
High CO_2 incorrectly classified as low

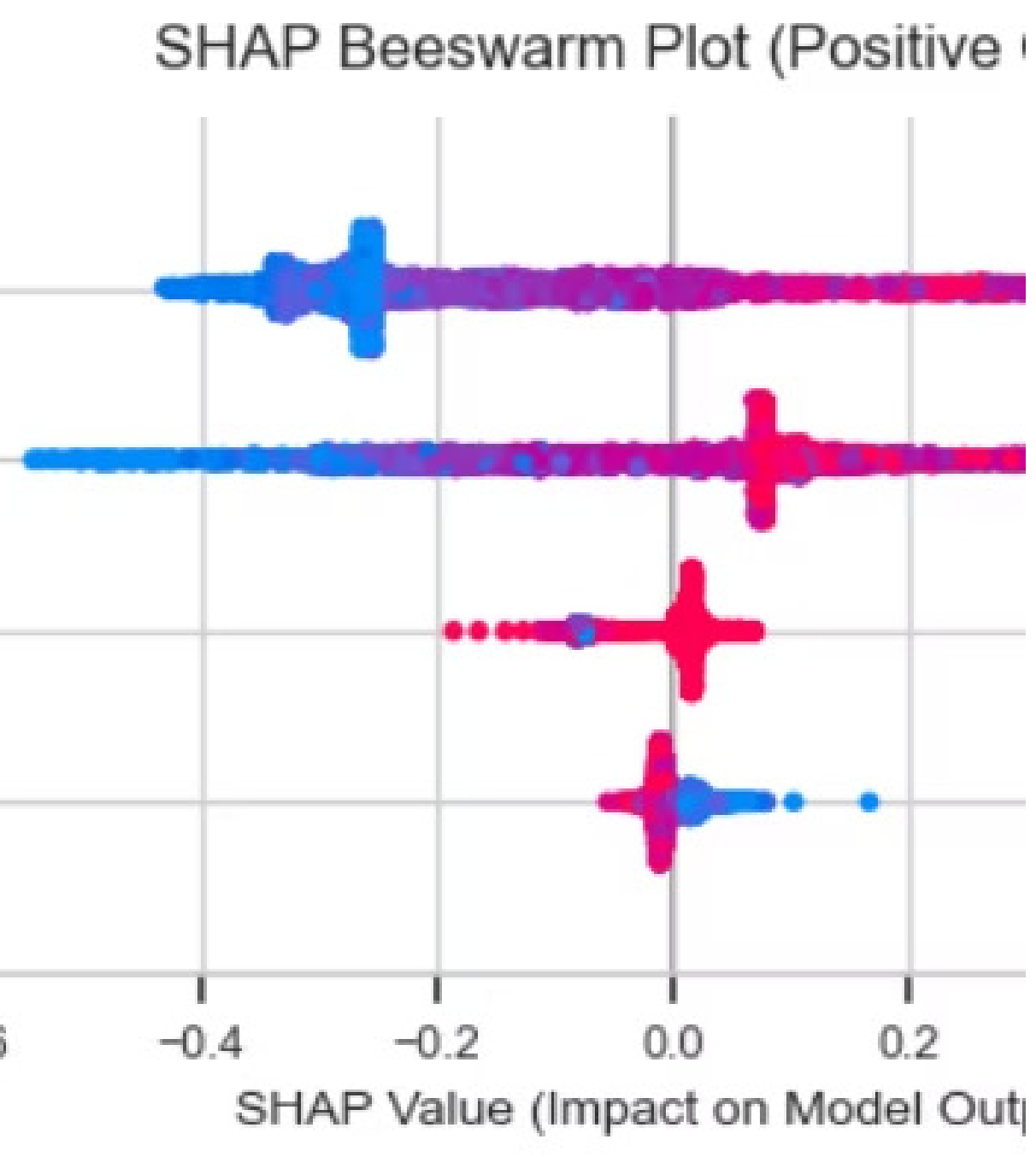
The model shows exceptional performance in identifying both high and low CO_2 conditions, with minimal misclassifications.

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_2 predictions



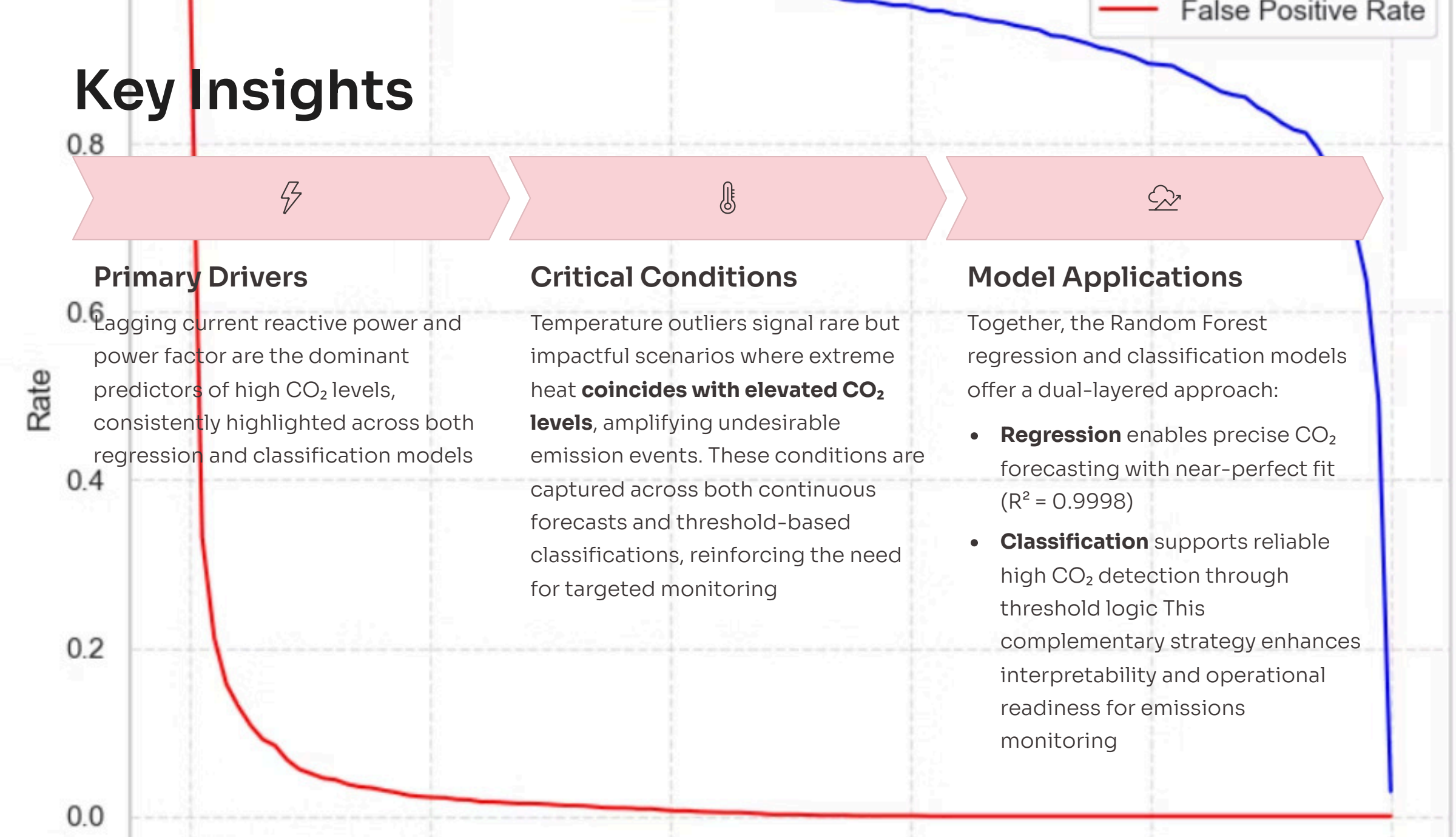


SHAP Dependence Analysis

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_2 risks
- Clear thresholds identified where feature values significantly shift predictions

Key Insights



Primary Drivers

Lagging current reactive power and power factor are the dominant predictors of high CO_2 levels, consistently highlighted across both regression and classification models

Critical Conditions

Temperature outliers signal rare but impactful scenarios where extreme heat **coincides with elevated CO_2 levels**, amplifying undesirable emission events. These conditions are captured across both continuous forecasts and threshold-based classifications, reinforcing the need for targeted monitoring

Model Applications

Together, the Random Forest regression and classification models offer a dual-layered approach:

- **Regression** enables precise CO_2 forecasting with near-perfect fit ($R^2 = 0.9998$)
- **Classification** supports reliable high CO_2 detection through threshold logic This complementary strategy enhances interpretability and operational readiness for emissions monitoring

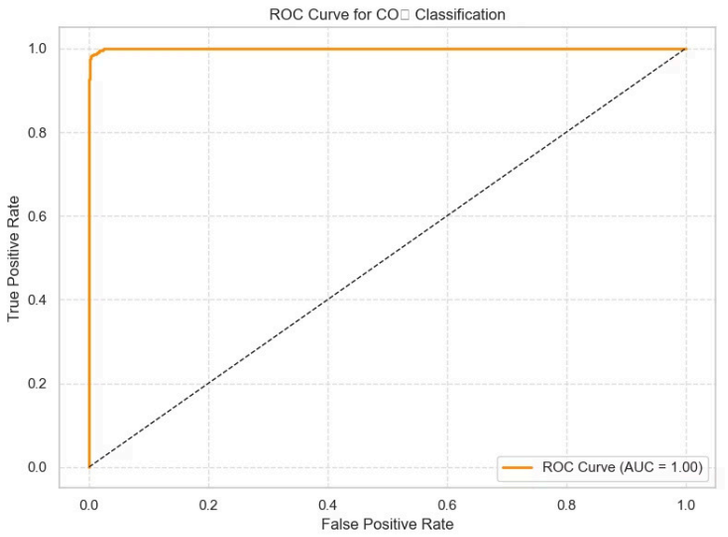
Overview and Windup

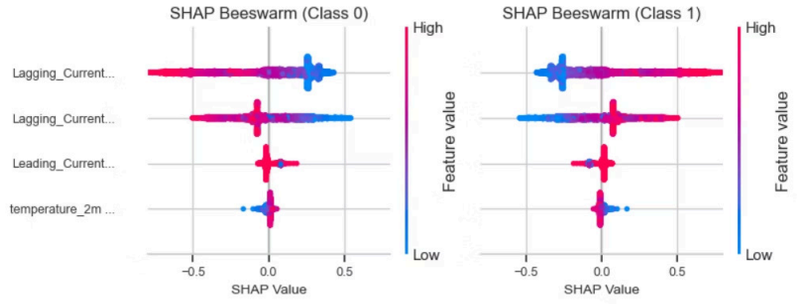
Conclusions

- Dual Random Forest models: regression and classification that successfully captured CO_2 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_2
- 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






```
X_test shape: (2811, 4)
shap_values shape: (2811, 4, 2)
Mean |SHAP values| (Class 0): [0.25977047 0.17145501 0.02603913 0.01329886]
Mean |SHAP values| (Class 1): [0.25977047 0.17145501 0.02603913 0.01329886]
```

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