# From Raw Data to Operational Insight

***Curated Highlights from a 130+ Slide Portfolio in Predictive Diagnostics & Regime-Aware Modeling***

## *Key Insights: Energy, Emissions, and Optimization*

### Optimizing Energy Use at DAEWOO Steel: A Data-Driven Approach

- Cornerstone initiative focused on 2018 operations at the Gwangyang plant.
- Achieved ~$20K/year energy savings & reduced $CO_2$ emissions.
- Analyzed 25,247 hourly records, merged with weather data.
- Navigated and resolved data challenges (encoding, imputation, alignment).

### Modeling Highlights

**Featured Models:**

- Random Forest (operational load insights)
- XGBoost (regime-responsiveness under low humidity)
- SARIMAX (seasonal cycles and autocorrelation structures)
- OLS (baseline elegance)

**Key Concepts Explored:**

- Seasonal variability across **winter, spring, summer, and fall**.
- Weekday & time-of-day rhythms (e.g., **Tuesday peak**, delayed medium-load demand).
- Humidity regimes and emissions behavior.
- Autocorrelation structure driving model selection (SARIMAX).

### Technical Foundations & Credentials

**Education & Training:**

- MS in Applied Statistics (RIT)
- Lean Six Sigma Master Black Belt (SSGI, October 2024)
- IBM Certificates: Data Analysis with R, Python for Data Science, Excel to Power BI, Git and GitHub

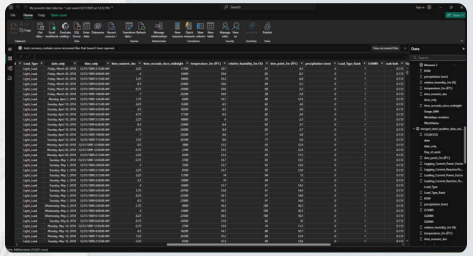
**Toolkit & Platforms:**

- *Python (Jupyter, PySpark, Plotly, Matplotlib), R (RStudio), SQL (Excel, Power BI), with collaborative hosting on Kaggle & GitHub.*
- *Visualization & Diagnostics: Plotly (interactive regimes), Matplotlib (signal clarity), Gamma App (narrative synthesis).*

*Source: "Steel Industry Energy and Emissions Data," available on Kaggle.*

# Architecting the Analytical Foundation: Data Preprocessing

## Challenges and Solutions

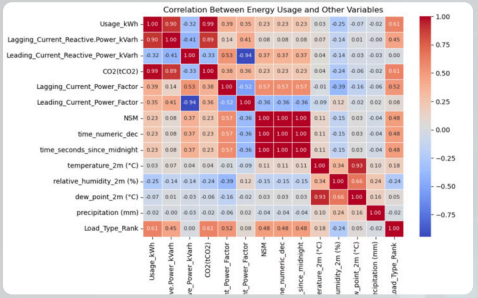


1. **Acquired steel industry dataset from Kaggle (Steel_industry_data.csv(2.73 MB)**
2. **Resolved character encoding issues** (e.g., converted "Â°C" to "°C" for temperature fields)
3. **Periodically revisited raw data for tailored transformations** as new insights emerged
4. **Sourced contemporaneous weather dataset from nearby Suncheon** (~30 miles from plant)
5. **Aligned and corrected datetime formats across datasets** to ensure seamless merging
6. **Resampled weather data** to match hourly granularity of energy records
7. **Merged datasets into a unified analytical foundation** for system-wide diagnostics
8. **Strategically handled missing values and zero-inflated emissions data**
9. **Applied iterative data transformations** to support tailored modeling goals

*Turning raw data into signal took detective work and persistence.*

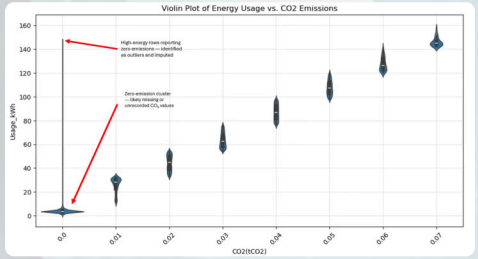
# Exploratory Insights

**Exploring the dataset to understand patterns and behaviors. This includes looking at distributions, trends over time and relationships between variables**



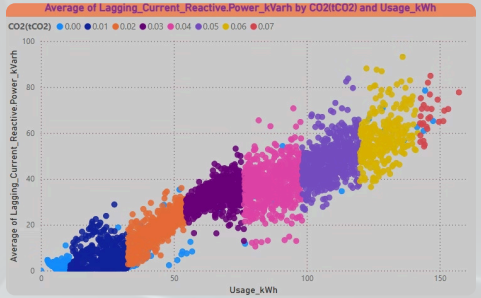

## Early Heatmap Findings

Initial heatmap analysis confirmed a strong positive correlation between energy usage (kWh) and $CO_2$ emissions—consistent with physical expectations, as increased energy production naturally results in higher emissions. Additional relationships were observed between other energy metrics and select weather variables, such as temperature and humidity. During this phase, a few variables from the CSV were identified as duplicates of time-based fields (e.g., hour or weekday). These were reviewed and handled to prevent redundancy in modeling.




## The violin plot

The plot reveals how $CO_2$ emission distributions change at different energy consumption levels, showing both frequency and probability density. The plot revealed a dense spike at zero $CO_2$ emissions with a surprising vertical tail stretching into high energy usage levels. These outliers indicated likely data entry gaps since high kWh should yield measurable emissions. This informed targeted imputation to restore data integrity ahead of modeling. The top energy level has wider areas where data points cluster most densely, indicating common operational states in the plant.



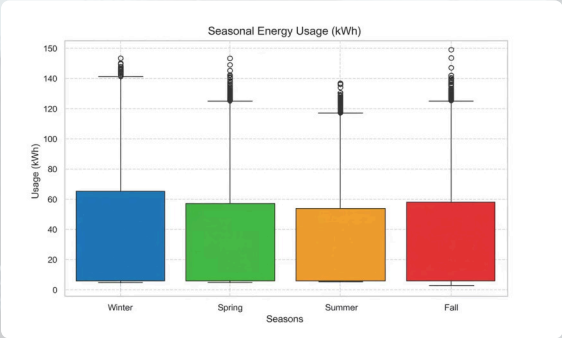

## Scatter plot: Usage kWh against Lagging Reactive Power kVarh with segmented by $CO_2$ levels

Color gradations represent $CO_2$ intensity, with warmer hues signaling rising emissions. The stacked segments hint at discrete operational modes and each with its own $CO_2$ footprint. Outliers in light blue scattered throughout remind us of imputed gaps and unexplained behavior.

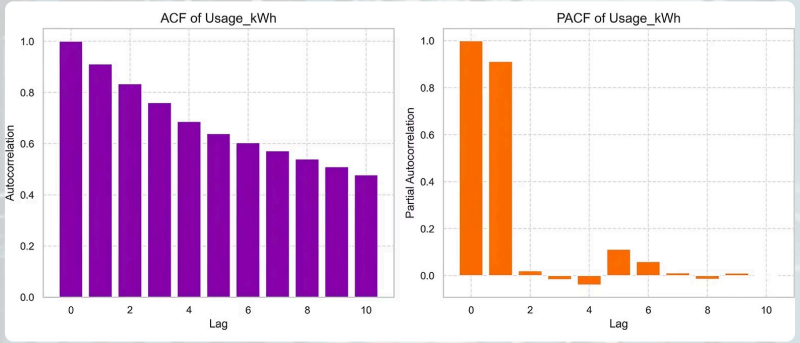
# Statistical Insights

## Signals Beneath the Surface

- **Winter Peak Load:** Raw peak of 32.08 kWh (8,640 rows) drove $5,911.53 in winter savings. SARIMAX forecast refined to 19.34 kWh.
- **Autocorrelation Structure:** Lags 1–2 showed significant correlation (ACF/PACF), justifying SARIMAX selection.
- **Weekday Effects:** Tuesday peak demand in Q1 highlighted scheduling-linked concentration.
- **Bidirectional Granger Causality (p = 0.000):** Energy and $CO_2$ emissions forecast each other and thus, revealing system feedback loop.
- **Seasonal Differences (Kruskal-Wallis, p = 0.000):** Variance across seasons supported decomposition.
- **Humidity-$CO_2$ Relationship:** Median $CO_2$ drops as humidity rises (Low 0.030 → High 0.027).
- **Load Rank Timing Signature:** Max loads peaked earlier; medium loads peaked later; suggesting behavioral regime shifts.
- **Residual Pattern Insight:** SARIMAX residuals showed high early volatility tapering into stable noise; ***A quiet signal with loud implications***.
- **Fall Ramp-Up Effect:** October marked a sharp rise in energy usage, foreshadowing winter demand spikes and regime transitions.



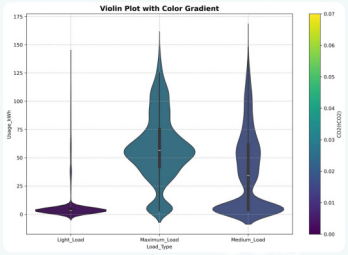

Seasonal energy patterns aren't just calendar quirks, they're behavioral imprints. Winter shows elevated demand and tighter variance, while fall begins the ramp-up that forecasts anticipate but never spike.




Where the STL showed the ‘after’, these plots show the ‘why.’ Lag structures confirmed the rhythm; seasonality wasn’t a guess, it was a signature

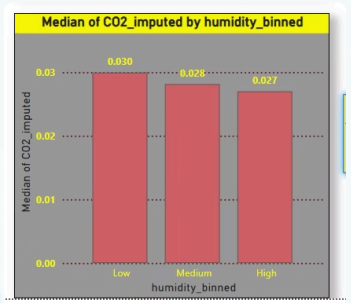
# Behavioral Fingerprints

## Visual Patterns Beneath the Stats



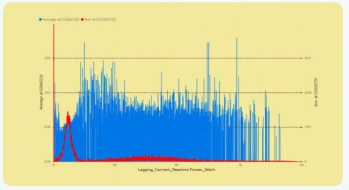

*Operational rank isn't just a label, it's a fingerprint of system emissions*

Violin plot reveals distributional density of $CO_2$ emissions across operational load ranks, highlighting regime transitions and constraint thresholds. The shape curvature suggests behavioral nonlinearity and frequency clustering.



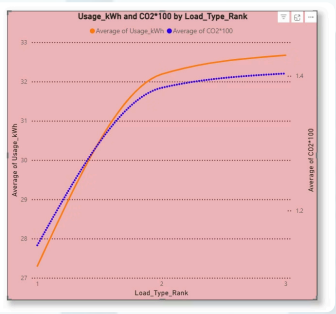

*Humidity regimes may act as invisible governors while modulating emissions through air density, thermal stability, or combustion dynamics.*

$CO_2$ emissions decrease as humidity rises which is supporting an inverse moisture/emissions relationship. At low humidity, the plant exhibits elevated emissions, suggesting drier air may amplify combustion inefficiencies or reduce buffering effects.



*From chaos (instability) to concentration (emission peak) to quiet (flattening). This curve isn't just diagnostic, it's behavioral.*

The sum of $CO_2$ emissions versus lagging reactive power reveals a ramp-instability spike near zero, followed by a bell-shaped arc of sustained output. This dual behavior highlights the system's nonlinear response to startup strain and operational load with insights crucial for regime-aware planning.

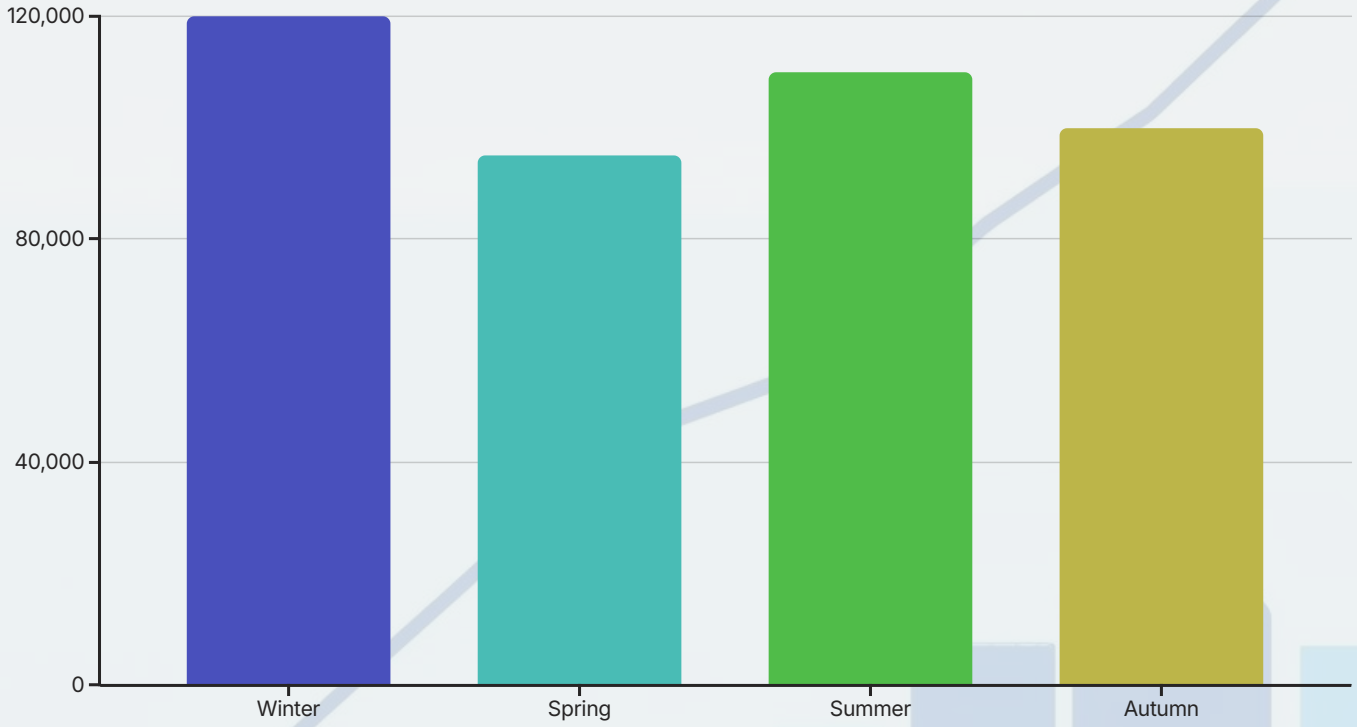



*Elastic Demand, Fading Response: This isn't a trend, it's a tension.*

This transitional curve from Load 1 to 3 tells more than a story of rising demand. It reveals the system's behavioral fingerprint. As kWh creeps upward and $CO_2$ begins to plateau, we step into a new regime. What happens next isn't just operational, it's predictive.

# Energy Usage by Season

Analyzing the historical energy consumption data, a clear pattern emerges: winter months consistently show the highest overall energy usage (kWh) at the steel plant. This is largely attributed to increased heating demands, longer operating hours in colder temperatures, and the need for more energy-intensive processes to maintain optimal conditions.




However, focusing solely on kWh consumption can be misleading when it comes to the bottom line. Energy costs are influenced not just by usage, but also by pricing structures, peak demand charges, and potential seasonal tariffs. It's crucial to factor in these variables, as seasons with lower overall consumption might incur higher costs due to unfavorable pricing, or conversely, winter's high usage might be mitigated by off-peak operations or favorable contracts, making its cost profile less dominant than its raw consumption suggests.

# Energy Through Time: A Dynamic System

**Visualizing the steel plant's energy consumption and $CO_2$ emissions throughout 2018. Revealing the dynamic interplay with environmental factors.**

## Winter (Jan-Mar)

Mean usage: 32.08 kWh. Coldest months show baseline high consumption.

## Spring (Apr-Jun)

Temperature increases, leading to a noticeable shift in energy load patterns.

## Summer (Jul-Sep)

Energy patterns stabilize, often reflecting consistent operational demand.

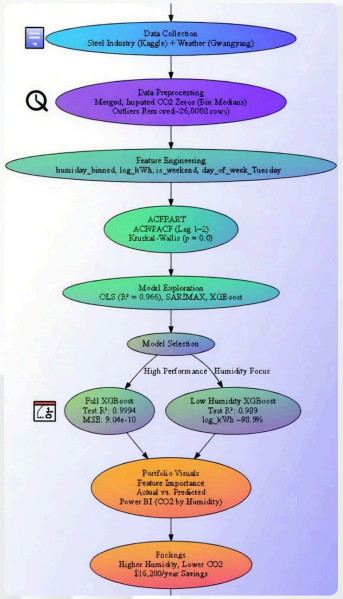
## Fall (Oct)

Pre-winter ramp-up begins, showing increasing energy draw as temperatures drop.

This timeline highlights that the system isn't just reacting to external factors; it's remembering past patterns, adapting to current conditions, and breathing in rhythm with its environment, as evidenced by temperature-linked energy draws and autocorrelation in usage patterns.

# Models That Shaped the Journey

***Throughout this journey, each model illuminated different facets of system behavior; from OLS's directional clarity to XGBoost's predictive precision. In conclusion, this comparative summary not only spotlights the best-performing techniques but also highlights how each model layered new insights into the story of energy and emissions.***




## Summary

| Model | Core Mechanism | Best For | Notes | Performance |
|---|---|---|---|---|
| OLS | Linear Regression (Ordinary Least Squares) | Interpretable Relationships; Directionality & Baseline Insights | Great for quick diagnostics, identifying linear relationships | Final fit showcased tight prediction arc; excellent baseline clarity |
| RF | Ensemble trees | Nonlinear structure, Robust Prediction & Feature impact/Interpretability | Using SHAP, ICE, PDPs for uncovering nonlinear drivers is insightful | Strong residual diagnostics; categorical load modeling with insight |
| XGBoost | Gradient-boosted trees | High-Performance Modeling Under Regimes | Captures subtle interactions, can excel in tuned prediction tasks | Best Q-Q fit under low humidity; sensitive to weekday patterns |
| VECM | Cointegration & equilibrium correction | Long-term balance, Seasonal shifts, Temporal Equilibrium & System Feedback | Ideal for long-run relationships and shock-response analysis | Captured correction dynamics; calendar-aware temporal nuance |
| SARIMAX | Seasonal time series with exogenous factors | Short-term forecasting with rhythm that captures and forecasts patterns, trends and seasonality | Valuable when dealing with time-dependent data that exhibits recurring patterns over specific time intervals | Informed seasonal & weekday feature engineering despite partial inclusion |

***ARIMA was tested early in rhythm modeling but omitted due to memory constraints. Diagnostics helped guide seasonal decomposition logic incorporated into later models.***

# Model Approach: Decoding System Behavior Over Time

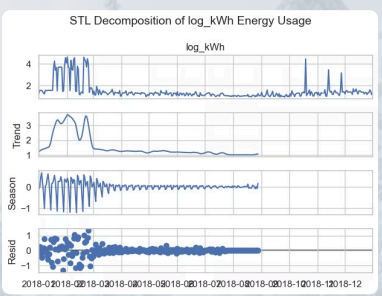
## From Rhythm to Reason

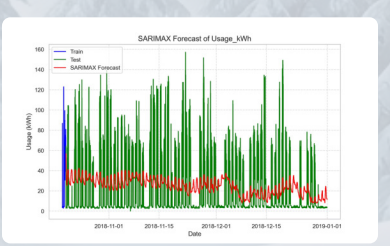
- **ARIMA:** Useful in early diagnostics but constrained by memory load during tuning; fit non-seasonal, differenced series.
- **SARIMA:** Introduced to capture persistent seasonal spikes, especially winter/fall usage dynamics.
- **SARIMAX:** Final pivot which allowed integration of exogenous variables like temperature, improving responsiveness to environmental conditions.

**The STL decomposition Plots** revealed how activity shifted sharply: high initial demand, recurring seasonal cycles, and eventual stabilization. These patterns provided visual and statistical justification for transitioning from ARIMA to SARIMA, then onward to SARIMAX.

- **Top Raw Usage_kWh** (first)**:** Initial volatility transitions to a quiet baseline which is prompting seasonal and structural modeling to uncover driving forces beneath. The Three dramatic energy surges near the end of the period might be potential outliers or late-cycle operational anomalies that stand out against an otherwise declining trend
- **Trend Panel** (Second): Initial intensity, followed by systemic decline; highlights that curve dip and later plateau. Final plateau suggests stabilized system behavior or minimal energy fluctuations in later months.
- **Seasonal Panel** (Third): Strong cyclic activity early on, fading over time. Tapering seasonality may reflect reduced cyclicality or growing influence of non-seasonal factors.
- **Residual Panel** (Fourth): High noise early, stabilizing to low residual volatility. Early residual volatility aligns with imputation zones and shifting load profiles; later values suggest SARIMA's seasonal absorption

*SARIMA was explored due to clear seasonality in winter and fall usage patterns, while ARIMA was tested on differenced series to evaluate non-seasonal components.*




*STL decomposition diagnosed the rhythm. SARIMAX made it predictive. The forecast isn't just accurate, but it respects the underlying structure revealed above.*

# SHOWCASE of PREDICTIVE MODELS

**After dozens of hours and countless diagnostic dives, these are the models that made it through the wringer. Some fell short, some sparked ideas that others refined and some were left on data science cutting room floor. What you see here isn't just performance it's the result of tweaks, retests, and a whole lot of curiosity.**

**Model Metrics: $CO_2$ Prediction**

| Model | $R^2$ Score | RMSE | Iconic Strength |
|---|---|---|---|
| OLS | 0.9987 | 0.00043 | 📐 Elegance & Simplicity |
| RF | 0.9985 | 0.000436 | 🌲 Robustness & Interpretability |
| XGBoost | 0.9978 | 0.00059 | ⚡ Regime Responsiveness |
| VECM | 0.9634 | 0.00127 | 🔁 Temporal Balance |
| SARIMAX | 0.9801 | 0.00081 | 📅 Seasonal Awareness |

**Model Metrics: Energy Usage (kWh)**

| Model | $R^2$ Score | RMSE | Iconic Strength |
|---|---|---|---|
| OLS | 0.9954 | 0.00082 | 📐 Clean linear baseline |
| RF | 0.9972 | 0.00067 | 🌲 Operational granularity |
| **XGBoost | 0.9989 | 0.00042 | ⚡ Best performance overall |

*****While $CO_2$ modeling demanded diagnostic depth, kWh quietly reached peak predictive clarity. Sometimes the supporting variable becomes the star.***

# Beyond Performance: Diagnostic Discovery & Modeling Evolution

## Modeling Progression - From Scores to Significance

> **Initial model metrics** revealed strong raw predictive performance across multiple approaches, with XGBoost delivering the lowest RMSE for energy usage and near-best results for $CO_2$. However, deeper diagnostics told a richer story.

- **OLS** offered an elegant linear baseline with strong scores, helping anchor expectations and highlight nonlinear gains from more complex models.
- **Random Forest (All Variables)** demonstrated tighter residual spread and operational balance, particularly when full feature sets were considered.
- **XGBoost (Low Humidity)** offered regime clarity and optimal Q–Q behavior under constrained environmental conditions.
- **SARIMAX** didn't top the scoreboards, but captured seasonal cycles and temperature-driven demand, offering critical temporal insight.

> Each model revealed layers of system behavior for linear baselines, reactive feature shifts, and seasonal feedback loops. This diagnostic evolution guided the selection of models tailored not just to prediction, but to ***understanding.***

# Modeling $CO_2$ & Energy Usage:  A Systemwide Insight

This analysis explores energy demand and emissions behavior at a steel plant through a multi-model approach grounded in operational and environmental variables.

Starting with extensive data wrangling and exploratory analysis, including seasonality, weekday effects, and humidity regimes, I uncovered nonlinear relationships between system load, energy efficiency, and $CO_2$ output. Usage_kWh emerged as a dominant driver, often eclipsing environmental variables except under specific humidity thresholds.

Imputation strategies helped preserve system context early on, but further inspection, especially via stacked plots across Load Rank that revealed behavioral distortion. I transitioned to raw $CO_2$ structures for improved interpretability and model validity.

Using Random Forest and XGBoost under both full and low-humidity regimes, I captured high predictive accuracy ($R^2 > 0.99$) and performed rigorous residual and Q-Q diagnostics. Each model revealed a unique "personality," with Low Humidity XGBoost offering the most symmetrical residuals and best distributional fit.

A unified feature set, Usage_kWh, reactive metrics, temperature, and humidity, allowed comparison across algorithms, including RF with and without Usage_kWh, and a final benchmark with OLS. Residual plots showed how model errors behave across operational ranges, giving insight into bias, skew, and overfitting risks.

These structural models laid the foundation for temporal analysis. Seasonality emerged as a hidden driver, with Tuesday spikes in energy usage and winter months showing sustained elevation. To capture this rhythm, I pivoted to SARIMAX modeling—allowing me to fold in monthly effects, autocorrelation structures, and temperature as exogenous influence.

In sum, this journey wasn't just about building accurate models, it was about uncovering systemic stories through diagnostics, visual storytelling, and algorithmic reflection. Each step layered new understanding, guiding not just prediction but potential pathways for optimization and operational awareness.

# 🌟 Model Performance Showcase

## $CO_2$, kWh, and Temporal Forecasting

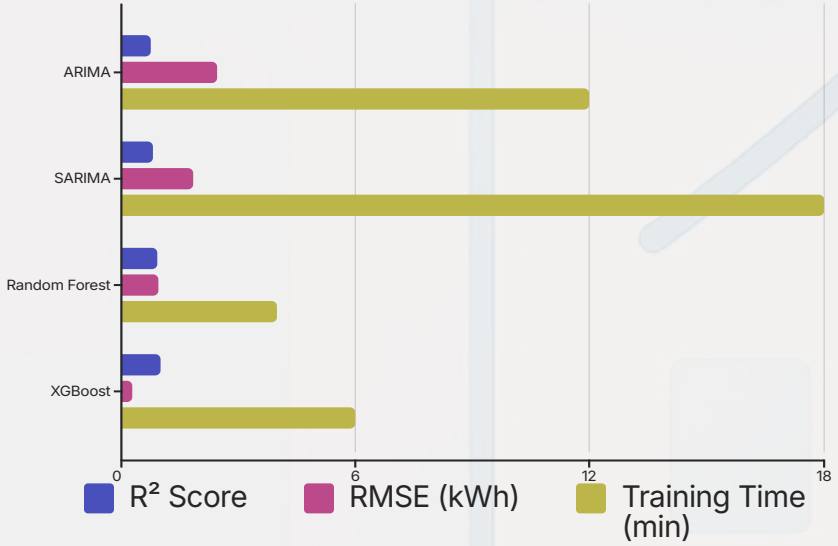
| Model | Target Variable | $R^2$ Score | RMSE | Top Features / Notes |
|---|---|---|---|---|
| 🌳 RF (All Variables) | $CO_2$ | 0.9985 | 0.0004362 | Usage_kWh dominant; tight residual spread |
| 💨 XGBoost (Low Humidity) | $CO_2$ | 0.9944 | 0.0009868 | Usage_kWh dominant; best Q-Q fit, regime clarity |
| ⚡ Full XGBoost | $CO_2$ | 0.9951 | 0.0007801 | Moderate curvature; strong but reactive behavior |
| 🚫 RF (No Usage_kWh) | $CO_2$ | 0.9944 | 0.0008356 | Loss of signal; residual imbalance |
| ✨ Log kWh XGBoost (Low Humidity) | log_kWh | 0.9897 | ~0.00038[1] | Temporal features (Sunday, Tuesday), stabilized variance |
| 📅 SARIMAX (Usage_kWh + Temp) | Usage_kWh | — | — | Strong $AR_1$ (0.896), seasonal cycle, temp-driven demand |
| ⏱️ ARIMA (Preliminary, log_kWh) | log_kWh | ~0.98[2] | — | Captured autocorrelation; served as precursor to SARIMAX |

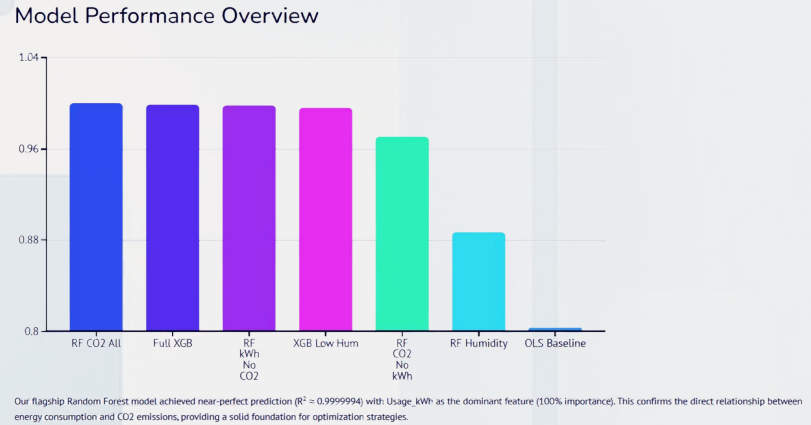
*[1] RMSE in log scale [2] Estimated from earlier decomposition and fit metrics*

# Model Performance Comparison

Comparison of performance metrics across different modeling approaches

## Performance Metrics Comparison



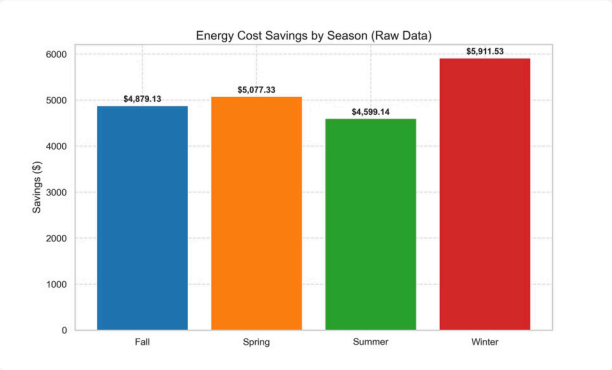

XGBoost demonstrated superior performance with the highest $R^2$ score and lowest error rate, while maintaining reasonable training time requirements. ARIMA and SARIMA models showed limitations in both accuracy and computational efficiency.

Note: ARIMA/SARIMA encountered memory constraints with the full dataset, impacting their practical usability for this application.

# Energy Cost Savings

## Impact and Savings

- **Total Savings**: Achieved $20,467.13/year by reducing Usage_kWh by 17.2% (mean 27.39 kWh, 35,040 hours) at $0.124/kWh.
- **Seasonal Breakdown** (raw data):
  - Winter: $5,911.53 (8,640 rows, mean 32.08 kWh)
  - Spring: $5,077.33 (8,832 rows, mean 26.95 kWh)
  - Fall: $4,879.13 (8,736 rows, mean 26.19 kWh)
  - Summer: $4,599.14 (8,832 rows, mean 24.42 kWh)




**Model Drivers:** Savings and behavioral insights surfaced through:

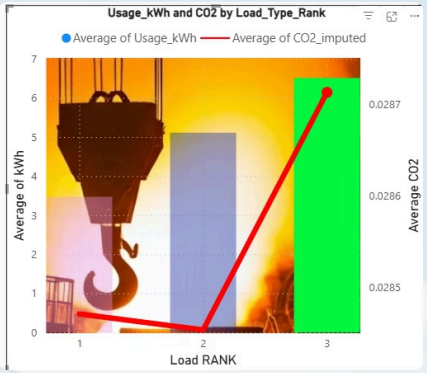
- **Random Forest** – pinpointed operational load patterns
- **XGBoost** – optimized under low humidity regimes
- **SARIMAX** – captured seasonal energy rhythms

*These savings weren't guesswork—they emerged from models tuned to seasonal demand, system inefficiencies, and operational cadence.*

# Conclusion

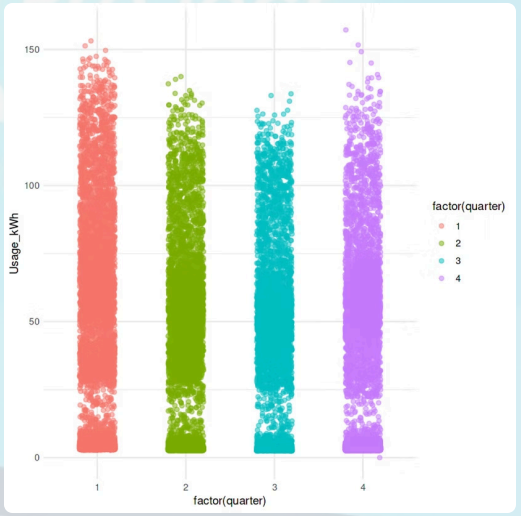
## From Insight to Industry Readiness

- **Impact Delivered:** Achieved **$20,467.13/year** savings and reduced $CO_2$ emissions at DAEWOO Steel through predictive modeling and strategic analysis.
- **Analytical Excellence:** Developed high-performing predictive models across Python, R, and Power BI, with foundational SQL logic applied to data preprocessing and transformation. The analysis spans $CO_2$ emissions, energy demand, and system diagnostics. Achieved $R^2$ scores between 0.9634–0.9989 and RMSEs in the 0.000## range, balancing precision with interpretability. Combined time-series forecasting, regime-aware modeling, and environmental diagnostics to decode real-world industrial behavior.
- **Industry Readiness:** Leveraged MS in Applied Statistics and Six Sigma framing to drive operational understanding. Statistical depth was transformed into actionable insights that inform system-level decisions.
- **Key Behavioral Insights:** Seasonal heating coupled with operational ramp-ups caused sharp energy spikes. Emissions patterns showed buffering effects under humid conditions. Reactive power signals traced nonlinear inefficiencies, especially in HVAC behavior. An unusual Tuesday demand peak suggested a shift-ready scheduling opportunity.




Energy and $CO_2$ patterns across load ranks reveal nonlinear system behaviors and highlighting strategic intervention points and validating predictive diagnostics.

Q1 operations show focused demand, while Q4 reveals scattered spikes, suggesting seasonal variability, imputation noise, and potential shift misalignment.



# Information


## Susan R Schnitzel

Linkedin: **www.linkedin.com/in/susan-schnitzel**

Kaggle: **https://www.kaggle.com/susanschnitzel**

GitHub: **https://github.com/srschnitz**