Regional Electricity Load Prediction Using Hybrid Machine Learning Models Under Limited Data Conditions

Kwesiga Kato Steven

2025/HD05/26350U :: 2500726350

November 2025

Challenge & Objective

Challenge:

- Accurate electricity forecasting is critical for grid stability and operational planning (UEDCL).
- Primary constraint: **limited historical data** and absence of external data like weather.

Objective:

- Evaluate a range of models—from simple to complex—for short-term load forecasting.
- Key question: Can smart data preparation outperform complex models?

Data Exploration: Patterns in Electricity Consumption

Dataset: Daily electricity consumption (kWh) for Kabalagala district, Jan 2021 – Nov 2025.

Observations:

- Clear upward trend in consumption.
- Strong weekly seasonality; distinct weekend patterns.
- Annual seasonality visible.

Conclusion: Time series has predictable patterns suitable for modeling.

Core Strategy: Feature Engineering

Approach: Transform patterns into features for machine learning.
Features:

- **Lag Features:** Previous day (`lag_1`), same day last week (`lag_7`), others (`lag_2`, `lag_3`, `lag_14`, `lag_30`).
- **Rolling Statistics:** Mean and std over 7, 14, 30-day windows.
- **Calendar Features:** Day of week, month, weekend flag, public holiday flag.

Model Evaluation: The Bake-Off

Models Tested:

- Baselines: Naive, Linear Regression, Random Forest
- Classical Time Series: SARIMAX
- Advanced ML: XGBoost
- Deep Learning: LSTM
- Hybrid: Weighted Ensemble

Validation: Temporal train-test split; last 90 days as hold-out.

Results: Surprising Winner

Model	RMSE	Key Takeaway
Linear Regression	1609.02	Winner
Hybrid Ensemble	1689.45	Strong performer
SARIMA	1879.07	Classical baseline
XGBoost	1944.94	Beaten by simpler model
LSTM	2457.21	Underperformed

Insight: Effective feature engineering enabled Linear Regression to outperform complex models.

Why Linear Regression Won: Residual Analysis

- Residuals over time: Randomly scattered around zero.
- Residual distribution: Normal, bell-shaped.
- Autocorrelation (ACF) plot: No significant spikes; temporal patterns captured.

Advanced Models Insights

- **XGBoost:** Feature importance validates strategy (lag_7, roll_std_7, lag_1 most important).
- **LSTM:** Limited data caused underperformance; deep learning requires more data.
- **Ensemble:** Second-best; confirms combining models stabilizes predictions.

Conclusion & Recommendation

Key Takeaway: In data-scarce environments, **feature engineering outperforms complex architectures**.

Recommendation:

- Deploy the feature-rich Linear Regression model at UEDCL.
- Most accurate, interpretable, and operationally practical.
- Guiding principle: **Invest in Features, Not Just Architectures.**

Q&A

Thank you.

Questions?