Final PROJECT REPORT

by

Syed Muhammad Umar

AI BOOTCAMP

Ghulam Ishaq Khan Institute of Engineering Sciences and Technology

August 2025

Report on Model Performance for Solar Energy Prediction

Introduction

In this project, I applied different machine learning and deep learning models — ARIMA, XGBoost, and LSTM — to forecast solar PV generation and load consumption based on the SkyElectric dataset. The goal was to predict power values at 10-minute intervals and achieve the best score on Kaggle using the provided competition platform.

Model Comparisons

1. ARIMA (Best Performing Model)

- Result: Achieved the best Kaggle score of 2019.
- Reason for Success:
    - ARIMA is well-suited for time series forecasting because it directly models temporal dependencies.
    - The dataset exhibited clear seasonality and trends, which ARIMA captured effectively.
    - With appropriate differencing and parameter tuning (p, d, q), ARIMA generalized well to unseen data.

2. XGBoost

- Result: Did not achieve competitive scores.
- Reason for Failure:
    - XGBoost works well with tabular structured data but struggles with sequential time-series dependencies unless heavy feature engineering is performed.
    - Even after adding time-based features (hour, day, month, lag values), the model overfitted and failed to generalize.
    - It could not capture long-term seasonality and temporal correlations.

3. LSTM

- Result: Did not yield fruitful results.
- Reason for Failure:
    - LSTM requires a large dataset and extensive hyperparameter tuning.
    - The dataset size was insufficient for deep learning to show advantages over ARIMA.
    - LSTMs are sensitive to scaling, and despite normalization, the model converged slowly and often overfit.
    - Computationally expensive compared to ARIMA.

Error Analysis (Test vs. ARIMA Submission)

I compared the ARIMA submission file with the actual test data. The following metrics were used:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)

Results:

- **PV Generation:**
  - **MAE = 170.4647364626048**
- **Load Consumption:**
  - **MAE = 160.9539301398877**

These results further confirmed that ARIMA provided the most accurate and consistent predictions among the tested models.

Feature Engineering & Preprocessing

- Timestamp-based features: Extracted hour, day, month, and weekday to capture cyclic behavior, historical and statical features.
- Normalization: Applied Min-Max scaling to bring PV generation and load values within the same range for LSTM and XGBoost.
- Missing Values: Handled anomalies and missing entries by interpolation.
- Stationarity: Differencing applied in ARIMA to stabilize mean and variance.

Conclusion

- ARIMA outperformed XGBoost and LSTM due to its natural ability to capture temporal patterns in time series data.
- XGBoost required more engineered features and still underperformed.
- LSTM was data-hungry and computationally heavy without producing better results.
- With a Kaggle score of 2019, ARIMA proved to be the most reliable model for this task.