ML Project: Hourly Electricity Load Prediction for Seven European Countries

Antoine Lenain, Benoit Lenain
Major: Energy and Sustainable Cities (EVD)
Course: Machine Learning Project 2025

November 15, 2025

1. Business Case

Problem: The stability of modern power grids relies on a precise, real-time balance between electricity generation and consumption. Grid operators must anticipate demand to prevent blackouts (under-production) or energy waste (over-production). Inaccurate forecasting is a significant operational and financial risk.

Objective: The primary objective of this project is to build and evaluate a machine learning model capable of accurately predicting the hourly electricity consumption (‘load$_{a}ctual‘)forsevenEuropeancount$

Field of Specialization: This objective is directly linked to our specialization in **Energy and Sustainable Cities (EVD)**. This project provides a practical understanding of grid management, the impact of renewable energy sources, and the high degree of interconnection between European energy networks. Our goal is not only to predict consumption but also to understand the key factors that drive it.

2. Description of Dataset and Source

Source: The data for this project was sourced from **Open Power System Data (OPSD)**, a public and reliable platform for energy-related time series data.

Dataset Description: The raw dataset (‘Adapted$_{D}ataset.csv‘)containshourlydataspanningfrom2015to2020,$
$theactualhourlyload, actualsolargeneration, andactualwindgenerationforoursevenselectedcountries.$

After a thorough data quality analysis, our final project scope included: **Austria (AT), Germany (DE), Spain (ES), France (FR), Denmark (DK), Greece (GR), and Romania (RO)**.

3. Data Exploration, Pre-processing, and Obstacles

Our data preparation was conducted in the ‘01$_{d}ata_{p}reparation.ipynb‘notebookandinvolvedovercomingseveralk$

- **Obstacle: Missing Values & Data Quality:**
  - **Analysis:** Our initial ‘Dataset analysis‘ revealed that several of our initially chosen countries (e.g., Italy, Belgium) suffered from severe data quality issues, with **>10-17% of data missing** for key renewable energy features.
  - **Solution:** We addressed this obstacle by performing a data-driven selection. We analyzed the entire dataset to identify a new cohort of 7 countries with high data integrity (<1% missing values). This led to our final list (AT, DE, ES, FR, DK, GR, RO). The few remaining gaps (<0.27%) were then filled using **time-based linear interpolation**, a method well-suited for time-series data.
- **Obstacle: Outliers:**
  - **Analysis:** Our visualization of the French load data (Notebook 1, Cell 6) revealed a single, massive, unrealistic data spike (approx. 160,000 MW). This was a clear ‘inconsistency‘.
  - **Solution:** We treated this outlier by replacing the erroneous value with ‘np.nan‘ and then re-applying our interpolation method. This fixed the data point without requiring us to delete the entire row.
- **Data Splitting (Addressing Overfitting):**
  - **Analysis:** For a time-series project, a random ‘train-test-split‘ is invalid as it causes "data leakage" (peeking into the future).
  - **Solution:** We implemented a strict, chronological split.
    - **Training Set:** 2015-2019 (43,706 hours)
    - **Test Set:** 2020 (6,576 hours)
- **Feature Engineering:**
  - To allow the models to understand temporal patterns, we extracted four new features from the datetime index: ‘hour‘, ‘day$_o f_w eek‘,‘month‘,and‘year‘$.

4. Formalization of the Problem

This project is a **supervised machine learning** task. We formalized the problem as a **multivariate time-series regression**.

- **Target (y):** We focused our modeling efforts on predicting a single target variable: the ‘FR$_l oad_a ctual_e ntsoe_t ransparency‘(Frenchhourlyload)$.

- **Features (X):** We used the remaining 24 columns as predictors. These included the load, solar, and wind data from the other 6 countries, the solar/wind data from France, and our 4 engineered time features.

5. Presentation of Models

We implemented three models to satisfy the project guidelines.

1. **Model 1: Linear Regression (Baseline):** This is the simplest standard model. It assumes a linear relationship between the 24 features and the target. Its purpose is to establish a "baseline" score which our advanced models must beat.

2. **Model 2: Random Forest Regressor (Advanced):** This is an advanced **Ensemble model**. It works by building hundreds of individual decision trees and averaging their predictions. It is highly effective at capturing complex, non-linear relationships that the Linear Regression model cannot.

3. **Model 3: Tuned Random Forest (Optimized):** This is the final model, where we used ‘RandomizedSearchCV‘ to perform **model’s hyperparameter’s tuning**. This process systematically tests different model configurations to find the optimal combination.

6. Comparison of Models Results

The models were trained on the 2015-2019 data and evaluated on the 2020 test data. The results are summarized below.

Model	**R^2 Score (R-squared)**	**MAE (Mean Absolute Error)**
1. Baseline (Linear Regression)	0.7732	4052.58 MW
2. Advanced (Random Forest)	0.8942	2635.88 MW
3. Tuned (Random Forest - Long)	**0.8971**	**2594.69 MW**

The **Linear Regression** baseline was already decent (R^2 0.77), but visualizations showed it failed to capture the volatility of peak and off-peak load.

The **Random Forest** provided a massive improvement, jumping to **R^2 0.8942** and reducing the average error by 1417 MW.

Our final **Tuned Model** (from the long search) provided a slight but measurable improvement, achieving the highest R^2 score (**0.8971**) and the lowest MAE. This confirms our model is robust and well-optimized.

7. Conclusion: How We Tacked Our Business Case

Our project successfully developed a highly accurate model for predicting hourly electricity load, achieving an **R^2 score of 89.7%**.

More importantly, by analyzing our best model (Notebook 2, Cell 8), we were able to achieve our secondary business objective: understanding what drives consumption. The **Top 10 Most Important Features** were:

1. ‘AT$_{l}oad_{a}ctual...$‘$(AustrianLoad)$

1. ‘month‘
2. ‘RO$_{l}oad_{a}ctual...$‘$(RomanianLoad)$
2. ‘ES$_{l}oad_{a}ctual...$‘$(SpanishLoad)$
2. ‘DE$_{l}oad_{a}ctual...$‘$(GermanLoad)$
2. ‘hour‘
3. ‘DK$_{l}oad_{a}ctual...$‘$(DanishLoad)$
3. ‘year‘
4. ‘FR$_{w}ind_{o}nshore...$‘$(FrenchWind)$
4. ‘DE$_{w}ind_{g}eneration...$‘$(GermanWind)$

This analysis provides two critical insights relevant to our **EVD major**:

1. **Seasonality and Interconnection are dominant:** The two most important features alone—Austrian Load (33.5%) and Month (31.9%)—account for over 65% of the model's decision-making.
2. **No Country is an Island:** The consumption in other European countries (AT, RO, ES, DE, DK) was a far better predictor of French load than France's own renewable production. This empirically proves the high level of grid interconnection and highlights that load patterns are shared across the continent.

8. References

- Open Power System Data (OPSD). *Data Package time_ series.* (2020). `https://data.open-power-system-data.org/time_series/2020-10-06`
- Scikit-learn developers. (2025). *Scikit-learn: Machine Learning in Python.* `https://scikit-learn.org`
- Lenain, A., & Lenain, B. (2025). *Project Step 1: Electricity Load Prediction.* (Internal Document).
- Mellouli, N. (2025). *Project Guidelines ML 2025 (Updates).* (Course Material).