Graphite Digital x Gridcare
Data Center Power Prediction Final Report

January 2025 - May 2025

Andy Xu, Sadhvi Narayanan, Aleicia Zhu, Anika Sharma, Felix Peng

1. Executive Summary

The goal of this project is to predict the total energy usage of data centers using free online datasets from SPGlobal, Aterio, and LBNL. These datasets were standardized to create a unified dataset for machine learning.

Key insights from our exploratory analysis revealed strong correlations between energy usage and features such as IT power allocation, IT space occupied, and total square footage. Climate was also shown to have a statistically significant effect on energy intensity, with regions like Warm Marine exhibiting higher average usage.

To handle missing data and improve model input quality, we applied multiple data imputation techniques, including K-Nearest Neighbors (KNN), Random Forest, Multiple Imputation by Chained Equations (MICE), and Neural Networks. We then trained several models—Decision Tree, Random Forest, XGBoost, and an Artificial Neural Network—using `TOTALPOWER` as the target variable.

The best-performing model achieved high predictive accuracy, with R-squared scores exceeding initial baselines and strong performance across our 5-fold cross-validation. These predictions can be used to estimate energy consumption for incomplete records, evaluate efficiency opportunities, and inform strategic decisions around sustainability and infrastructure planning.

2. Data Sources and Standardization

Sources Used:

- SPGlobal
- LBNL Building Performance Database
- Aterio US Data Center Power Demand Dataset

Rationale for Source Selection:
We focused on SPGlobal, LBNL, and Aterio because they offered the most structured, complete, and relevant data for modeling data center energy usage at scale.

Source	Sample Size	Key Features	Reasons for Inclusion
SPGlobal	~600 U.S. data centers	Detailed schema including power feeds, IT space, climate, square footage	Chosen as schema baseline/template due to completeness and coverage.
LBNL Building Performance Database	274 data centers	Climate zones, square footage, energy use metrics	Provided regionally coded, statistically robust data
Aterio US Data Center Power Demand Dataset	115 data centers	Measured TOTALPOWER in megawatts, facility metadata	Offered actual power usage, valuable for target modeling

Why Other Sources Were Excluded:

Source	Sample Size	Key Features	Reasons for Exclusion
EIA CBECS (Energy Information Administration)	N/A (data centers bundled in "other")	Building-level energy consumption across commercial types	Lacked specific data center identifiers; too generic
Individual Utility Records & Spec Sheets	Highly variable	Potentially detailed operational records	Too fragmented, inconsistent, and labor-intensive to collect at scale

3. Exploratory Data Analysis

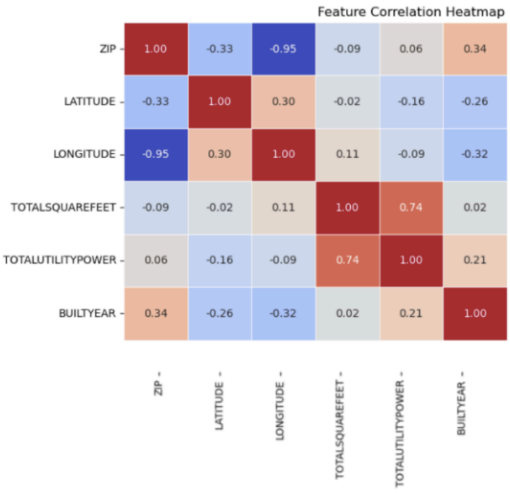
Correlation Analysis

We conducted feature correlation analysis separately and in aggregate on the SPGlobal, LBNL, and Aterio datasets to understand the drivers of energy usage.

- Key correlated features across datasets included:
  - Total Square Footage
  - Total IT Space Occupied
  - Total Number of Racks
  - Year Built
  - Zip Code/Location
- Watts Per Square Foot was more consistent between data centers, likely because of similar optimization strategies and standards.
- In SPGlobal, these features showed strong positive correlations with `TOTALPOWER`, with coefficients as high as 0.80.
- Climate effects from LBNL were also analyzed using ANOVA, revealing significantly higher energy use in Warm Marine climates and lower in Cool Humid regions.
- Combined, our dataset contained 202 unique data centers in the US and internationally.

Visualizations:

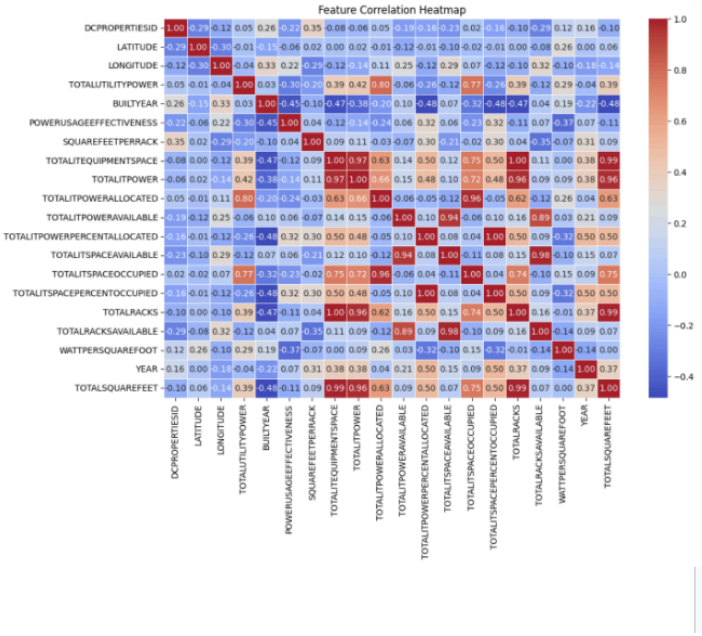
Aterio Dataset



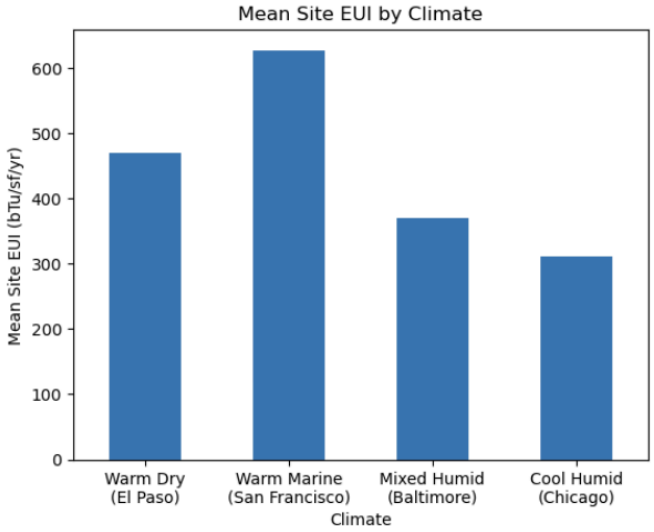
- Plotted a correlation matrix on the features
- Total **square feet** and **Year Built** seem to be the most related to the total utility power of a data centre
- Other notable correlations:
  - **Built year and Zip code:** need to confirm this correlation by analyzing the data case by case

SPGlobal

- Correlations with Total Utility Power:
  - Total IT Power Allocated (0.805)
  - Total IT Space Occupied (0.773)
  - Total IT Power (0.425)
- Least correlated features:
  - Power Usage Effectiveness (-0.304)
  - Total IT Space Percent Occupied (-0.256)
- Year built didn't seem to correlate (0.04), but square feet/total racks was somewhat correlated



LBNL - Energy use intensity by climate



- Ran ANOVA and post-hoc t-tests
- Warm Marine significantly higher than pooled mean
- Cool Humid significantly lower

*Cool Dry omitted for limited data

4. Imputation Strategy

To address missing values in the combined dataset, we implemented multiple imputation methods.

Methods Used:

- K-Nearest Neighbors (KNN)
- Random Forest
- Multiple Imputation by Chained Equations (MICE)
- Neural Network-based imputation

Model	RMSE	MAPE	RMSLE	R2
mice_df	9763.1083984375	3656493056000.0	3.2686712741851807	0.2939906120300293
random_forest_df	4035.999267578125	1630815700.0	1.9057868719100952	0.7349888980388641
knn_df	6187.7080078125	1201597030400.0	2.6994035243988037	0.37709540128707886
neural_df	4411.79052734375	1692063400.0	0.9255340099334717	0.6833411455154419

5. Model Comparison

Models Evaluated:

- Decision Tree
- Random Forest
- Neural Network

Evaluation Metrics:

- Accuracy_score from sklearn.metrics
- Precision and Recall
- f1-score

These metrics provided a comprehensive view of both absolute and relative prediction accuracy for the classification problem.

Validation Approach:

- Train-Test Split: Initial performance was evaluated using an 80/20 train-test split
- Cross-Validation: 5-fold cross-validation was used to assess model stability and reduce variance in performance estimates.

6. Modeling Pipeline

Target Variable:

- `POWERBUCKETS (in 3 MW buckets)`

Feature Engineering:

- Where `TOTALPOWER` was missing, it was estimated using:
  `TOTALPOWER = (Total Square Footage × Kilowatts per Rack) ÷ 1000`

Preprocessing Steps:

- StandardScaler for normalizing numerical features
- OneHotEncoder for categorical variables (e.g., climate, market, state)

Input Features:

```
DCPROPERTIESID, YEAR, ZIP, CITY, STATE, CLIMATE, BUILTYEAR,
DATACENTERNAME, STREETADDRESS, LATITUDE, LONGITUDE,
NUMBEROFPOWERFEEDS, UPSREDUNDANCY, POWERUSAGEEFFECTIVENESS,
SQUAREFEETPERRACK, TOTALITPOWERAVAILABLE, TOTALITSPACEAVAILABLE,
TOTALITSPACEPERCENTOCCUPIED, TOTALRACKSAVAILABLE, SOURCE,
TOT_DATACENTER_SPACE_SQFT, COUNTRY, COUNTRYDIVISIONCODE,
COUNTRYID, FOREIGNPROVINCE, FOREIGNZIPCODE, GENERATORREDUNDANCY,
GEOGRAPHICREGION, ID, MARKET, MSAID, MSANAME, TENANCY, UPDDATE,
UPTIMECONSTCERTIFICATION, UPTIMEDESIGNDOCCERTIFICATION,
UPTIMEOPSUSTCERTIFICATION, YEARRETROFITTED, CURSOR,
KILOWATTSPERRACK, NETRACKS, OPERATIONALSTATUS, QUARTER,
QUARTERENDDATE
```

Output Feature:

- `POWERBUCKETS (in 3000 kW buckets)`

7. Conclusion and Recommendations

Best-Performing Model

Based on the summary table results, the random forest model outperformed neural networks and decision trees on the accuracy and f1 scores. This remained true even after we added measures to prevent overfitting, such as reducing the tree depth on random forest. The full evaluation metrics for the random forest model are shown below:

Metric	Precision	Recall	F1-Score	Support
Accuracy			**0.79**	**33**
Macro Avg	0.64	0.66	0.61	33
Weighted Avg	0.76	0.79	0.76	33

Bucket	Precision	Recall	F1-Score	Support
0–3000	0.88	1.00	0.94	15
15000–18000	1.00	1.00	1.00	1
3000–6000	0.78	0.78	0.78	9
30000–33000	0.00	0.00	0.00	2
36000–39000	0.33	1.00	0.50	1
6000–9000	1.00	0.50	0.67	2
9000–12000	0.50	0.33	0.40	3

Random forest is suitable for the high-dimensionality of our data set and is better able to discriminate the relevant features. Hence, our modeling pipeline employs random forest for prediction. We predicted power using a power bucket classification rather than a single power measurement to allow for an approximate power range measurement in 3 MW buckets.

Potential Improvements/Further Data Collection

- Expand coverage of missing fields (e.g., wattage per rack, climate zone) to reduce reliance on imputation.
- Gather time-series data (quarterly or monthly energy usage) for dynamic modeling.
- Include more granular infrastructure metrics (e.g., cooling type, server density).

8. Usage instructions

The main model pipeline can be found on the main branch in the github repository at the following path: *final_model_experimentation/model_pipeline.ipynb*

1) The model requires an input JSON with particular parameters. An example JSON is given in *final_model_experimentation/features.json*. When data is not available, the pipeline expects "None" as the value for that parameter.
2) The jupyter notebook should run as a standalone file apart from the following manual intervention:
    a) After the imputation of the dataset there is a code segment to consider which features are highly correlated with one another. Immediately below this there is a list named *features_to_drop.* The goal here is to drop features which are unlikely to be available in a real-world setting before training the model — through which power can be inferred directly. For example, in our dataset these were features like: TOTALUTILITYPOWER and WATTPERSQUAREFOOT. These may vary from dataset to dataset. Hence if using another dataset, one should look over the features to determine which ones to include and remove before training the model.
3) The trained model saves to a pickle file called *model.pkl* in the same directory.

9. Appendix

Thank you to Arushi, Adam, Aashish, and everyone at Gridcare who supported us during the course of this project!

3/4/2025 GG Digital x Gridcare Initial Data Sources Summary
4/2/2025 GG Digital x Gridcare Initial Data Analysis Summary
5/7/2025 GG Digital x Gridcare Feature Selection and Model Refinements

Github link: https://github.com/GraphiteGroupDigital/gridcare.git