Analyst Report: Predictive Engine for Building Energy Efficiency

M4U1 – Data Analytics for AECO

---

1. Executive Summary

This report presents a Predictive Engine built to demonstrate how an AECO firm can leverage historical building data to predict future project outcomes — specifically, a building's **Heating Load** (energy demand in kWh/m²). Two machine learning models were trained and compared: **Linear Regression** and **Random Forest Regressor**. After evaluation and hyperparameter tuning, the **Random Forest model** emerged as the clear winner with an R^2 score of **0.9977** and an RMSE of only **0.49 kWh/m²**.

---

2. Dataset Selection

Dataset: UCI Energy Efficiency Dataset

- **Source:** UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/242/energy+efficiency)
- **Authors:** A. Tsanas & A. Xifara (2012), published in *Energy and Buildings*, Vol. 49
- **Size:** 768 building configurations × 10 variables
- **Format:** Excel/CSV (no missing values)

Features (8 input variables):

Variable	Description	Unit
X1	Relative Compactness	Ratio

Variable	Description	Unit
X2	Surface Area	m^2
X3	Wall Area	m^2
X4	Roof Area	m^2
X5	Overall Height	metres
X6	Orientation	Categorical (N/E/S/W)
X7	Glazing Area	Ratio
X8	Glazing Area Distribution	Categorical

Target Variable:

- **Y1 — Heating Load** (kWh/m^2): The energy required to heat the building.

Why is this dataset relevant to AECO?

1. **Building Performance is Core to AECO:** Architecture, Engineering, Construction, and Operations firms routinely collect data on building geometry, materials, and envelope properties. This dataset mirrors the type of spreadsheet data an AECO firm would have from past projects.
2. **Energy Efficiency is a Business Priority:** With LEED certification, green building mandates, and rising energy costs, predicting heating/cooling loads during the design phase saves money and improves sustainability outcomes.
3. **Pre-Construction Decision Support:** By predicting energy loads from design parameters, the firm can optimise HVAC sizing, reduce over-engineering, and provide clients with accurate energy cost estimates before construction begins.

---

3. Algorithm Selection

Model A — Linear Regression

Why chosen: Linear Regression is the simplest supervised learning algorithm. It serves as an excellent baseline because:

- It is highly interpretable (coefficients show the direction and magnitude of each feature's effect)
- It trains instantly and requires no hyperparameters
- If the relationship between features and target is approximately linear, it performs well
- Business stakeholders can easily understand its outputs

Model B — Random Forest Regressor

Why chosen: Random Forest is an ensemble method that builds hundreds of decision trees and averages their predictions. It was selected because:

- It captures **non-linear relationships** and **feature interactions** that Linear Regression misses
- It is robust to outliers and does not require feature scaling
- It provides built-in **feature importance** rankings
- It is widely used in industry for tabular regression problems

---

4. Results

Initial Comparison

Metric	Linear Regression	Random Forest (Default)
R^2 Score	0.9122	0.9977
RMSE (kWh/m²)	3.03	0.49
MAE (kWh/m²)	2.18	0.35

Interpretation: Linear Regression explains 91% of the variance — a respectable result. However, Random Forest explains 99.8% of the variance with 6× lower prediction error. The remaining 8.8% of variance that LR misses is due to non-linear interactions between building geometry variables.

Hyperparameter Optimisation (Random Forest)

GridSearchCV with 5-fold cross-validation was used to tune:

- `n_estimators`: [100, 200]
- `max_depth`: [15, None]
- `min_samples_split`: [2, 5]
- `min_samples_leaf`: [1, 2]

Best Parameters Found:

- n_estimators = 200
- max_depth = 15
- min_samples_split = 2
- min_samples_leaf = 1

Final Comparison (After Tuning)

Metric	Linear Regression	RF (Default)	RF (Tuned)
R^2 Score	0.9122	0.9977	0.9977
RMSE (kWh/m²)	3.03	0.49	0.49
MAE (kWh/m²)	2.18	0.35	0.35

The default Random Forest was already near-optimal for this dataset. Tuning confirmed the configuration and slightly refined the model depth for better generalisation.

---

5. Feature Importance

The tuned Random Forest identified the following feature importance ranking:

Rank	Feature	Importance
1	Relative Compactness	39.8%
2	Surface Area	21.0%
3	Overall Height	14.3%
4	Roof Area	12.4%
5	Glazing Area	7.8%
6	Wall Area	3.4%
7	Glazing Distribution	1.2%
8	Orientation	0.1%

Key Insight: Building compactness (the ratio of volume to surface area) is by far the most important predictor of heating load. This aligns with thermodynamic principles — compact buildings lose less heat. Orientation has almost no effect, suggesting heating demand is dominated by envelope geometry rather than solar exposure.

---

6. Business Recommendation

Which model should the company use?

Random Forest Regressor is the recommended model for production deployment.

Reasons:

1. **Near-perfect accuracy ($R^2 = 0.998$)** — Predictions deviate by only ~0.5 kWh/m² from actual values

2. **Captures real-world complexity** — Building energy performance involves non-linear interactions that LR cannot model
3. **Feature importance provides actionable insights** — Designers can focus on compactness and surface area to optimise energy performance
4. **Scalable** — The model can be retrained as new project data is collected

Practical Applications for the Firm:

- **Pre-construction energy estimation** for client proposals
- **Design optimisation** by simulating different building geometries
- **LEED/green certification support** with quantitative energy predictions
- **Cost forecasting** by linking predicted heating load to energy prices

---

7. Reproducibility

All code is provided in the accompanying Jupyter Notebook (`M4U1_Predictive_Engine.ipynb`). The analysis uses:

- Python 3.x with pandas, numpy, scikit-learn, matplotlib, seaborn
- Data is loaded directly from the UCI repository (no local files needed)
- Random state = 42 ensures identical results on re-run
- 80/20 train-test split

Report prepared as part of the M4U1 Data Analytics assignment for the Master's in AI for Architecture & Construction programme.