Urban Heat Island Prediction and Mitigation

Yifan Sun

Abstract

This project presents a ML-based pipeline for Urban Heat Island (UHI) analysis. The first component delivers comprehensive exploratory data analysis covering correlations, distributions, spatial quicklooks, and a dedicated geological module that examines elevation effects, estimates lapse rates, summarizes slope/aspect patterns, and maps residual spatial structure. The second component implements a full machine learning workflow, including feature engineering, preprocessing, model selection with randomized hyperparameter tuning, and side-by-side comparisons across linear, kernel, and ensemble methods. Evaluation on a held-out test set shows tightly clustered errors; tree ensembles and regularized linear models perform similarly, while slightly negative R^2 indicates a challenging target best addressed by richer features and geo-aware validation.

1 Introduction

Urban Heat Islands (UHIs) arise from the interplay of land cover, urban morphology, energy use, and atmospheric conditions in built environments, making rigorous data preparation and diagnostics essential for reliable inference and effective mitigation planning. This work delivers a two-part, self-contained pipeline: an EDA and geological analysis component that produces end-to-end diagnostics with publication-ready figures—covering elevation deciles, lapse-rate estimation, and local slope/aspect effects—and a modeling and tuning component that performs feature engineering, column-aware preprocessing, randomized-search model selection, and consistent evaluation with interpretable diagnostics. The workflow assumes a tabular dataset with a temperature target (e.g., `Temperature_degC`) and common covariates such as elevation, greenness, wind speed, humidity, rainfall, energy consumption, and population density, with optional latitude and longitude fields to enable spatial quicklooks and terrain-aware modules.

2 Methodology

2.1 Data Ingestion and Standardization

The workflow begins by harmonizing column names using a consistent renaming scheme (for example, converting temperature fields with units into a canonical `Temperature_degC`). After standardization, it computes descriptive statistics for both numeric and categorical variables and summarizes missingness patterns through aggregated tables and visual diagnostics, including a per-column bar chart and a row–column heatmap (restricted to the first N rows for readability). These steps ensure that downstream analyses operate on clean, uniformly labeled features with transparent data quality profiles.

2.2 Exploratory Data Analysis (EDA)

The exploratory stage provides a comprehensive set of diagnostics designed to surface linear and non-linear structure, distributional quirks, and spatial patterns relevant to urban heat. It produces Pearson and Spearman correlation heatmaps and a ranked table of the strongest pairwise associations; per-feature distribution checks combining histograms with kernel density overlays and normal Q–Q plots; group comparisons for numeric variables stratified by land-cover categories; and bivariate hexbin plots relating temperature to key drivers with fitted trend lines. When geographic coordinates are available, a spatial quicklook maps temperature over longitude and latitude, and a compact scatter-matrix visualizes the features most strongly related to the target.

2.3 Geological Analysis

To incorporate terrain effects, the geological module augments the EDA with elevation-aware and topographic diagnostics. It profiles temperature across elevation deciles and visualizes the relationship between temperature and elevation using density-aware bivariate plots; estimates an atmospheric lapse rate by regressing temperature on elevation with optional latitude/longitude controls and inspects spatial residuals to reveal remaining structure; and derives local slope and aspect via neighborhood plane fitting, reporting slope distributions, temperature–slope relationships, aspect polar histograms, and temperature aggregated by aspect sectors. Together, these views isolate orographic influences that can confound or amplify urban heat signals.

2.4 Feature Engineering and Preprocessing

Feature engineering expands the predictor set with domain-informed transforms and interactions, including logarithmic stabilizations for skewed variables (e.g., population density, energy use, income), ratio constructs that capture greenness versus imperviousness, and interaction terms reflecting density–greenness, energy–air-quality, and elevation–humidity couplings. After a single pass to materialize these derived features, the pipeline enumerates numeric and categorical columns and applies a column-wise preprocessing strategy: median imputation and standardization for numeric variables, and most-frequent imputation with one-hot encoding for categoricals. This design preserves column provenance for interpretability, avoids non-picklable constructs, and yields a consistent, model-agnostic feature space.

2.5 Model Selection and Hyperparameter Tuning

Modeling proceeds with a diverse set of estimators spanning linear regularized methods (Ridge, Lasso, Elastic Net), kernel and instance-based learners (RBF-SVR, KNN), and tree ensembles (Random Forest, Extra Trees, Gradient Boosting), with the option to include gradient-boosted trees from external libraries when available. Each model is embedded in a unified pipeline comprising feature engineering, preprocessing, and the estimator itself, and is tuned via randomized search with k-fold cross-validation using negative MAE as the selection metric. The workflow ranks models by held-out performance, persists the strongest configuration for each family, and identifies a global best model for interpretation and deployment through permutation importance, partial-dependence visualizations (when supported), learning-curve diagnostics, and test-set prediction summaries.

3 Evaluation

The modeling script evaluates performance using a standard 80/20 train–test split and summarizes results with four complementary metrics: the mean absolute error (MAE), which captures the average absolute deviation in degrees Celsius; the root mean squared error (RMSE), which places greater weight on larger residuals; the mean absolute percentage error (MAPE), a scale-free measure of relative error; and the coefficient of determination R^2, which quantifies the proportion of variance in observed temperatures explained by the model.

4 Results and Discussion

4.1 Model Comparison

Table 1 reports model performance from an 80/20 train–test split using four complementary metrics (MAE, RMSE, MAPE, and R^2), with models ranked by MAE. Errors are tightly clustered: MAE spans roughly 6.13–6.28 °C and RMSE 7.02–7.15 °C. The strongest performers by MAE are Random Forest (6.126; RMSE 7.045), Gradient Boosting (6.140; 7.034), and the Lasso/Elastic Net pair (both 6.151; 7.023), indicating that additively separable effects captured by feature engineering explain much of the signal and that nonlinear gains are modest under the current covariates. SVR and KNN trail slightly (MAE 6.214 and 6.277), suggesting sensitivity to scaling/noise or limited benefit from kernelized locality. MAPE sits near 35% across models, consistent with scale heterogeneity or residual heteroscedasticity, and R^2 values are mildly negative (-0.064 to -0.028), implying a challenging target with notable irreducible variability or distribution shift on this split. These findings motivate (i) enriched predictors (impervious subclasses, building height/plan density, sky-view factor), (ii) geo-aware or group-wise cross-validation to better match train/test distributions, and (iii) broader model/tuning exploration (e.g., histogram-based gradient boosting, CatBoost, stacking) while guarding against leakage; given the competitiveness of regularized linear baselines, careful feature curation and variance reduction (e.g., denoised greenness, seasonal controls, reanalysis features) may yield larger gains than added algorithmic complexity alone.

Table 1: Comparing Models (lower MAE/RMSE/MAPE% is better; higher R^2 is better).

Model	**MAE**	**RMSE**	**MAPE%**	$\mathbf{R^2}$
RandomForest	6.1260	7.0453	35.1873	−0.0343
GBDT	6.1402	7.0339	35.1113	−0.0309
Lasso	6.1507	7.0233	35.1708	−0.0278
ElasticNet	6.1507	7.0233	35.1708	−0.0278
Ridge	6.1570	7.0555	35.2877	−0.0373
ExtraTrees	6.1587	7.0468	35.2356	−0.0347
SVR	6.2139	7.1258	36.0564	−0.0580
KNN	6.2771	7.1465	35.6188	−0.0642

4.2 Permutation Importance

Permutation importance highlights the features that most strongly reduce prediction error in the globally best-performing model. The top drivers include `Elevation_m`, `Energy_Consumption_kWh`, `Humidity`, and spatial coordinates (`Longitude`, `Latitude`), alongside interaction and domain-derived features such as `Greenness_x_Density`, `Energy_x_AQI`, and `Impervious_ratio`. This pattern is consistent with the combined influence of terrain, energy demand, moisture conditions, urban form, and air quality on near-surface urban temperature. Figure 1 embeds the produced bar chart (decrease in MAE when each feature is permuted), illustrating relative contributions; note that the absolute bar lengths are directly comparable within this model family and dataset.

4.3 Learning Curve

The learning-curve diagnostic in Figure 2 plots training and cross-validated MAE as a function of training set size. The persistent gap between training and validation error suggests moderate variance, while the downward trend in validation error with more samples indicates headroom for performance gains via additional labeled data or targeted data augmentation (e.g., seasonal

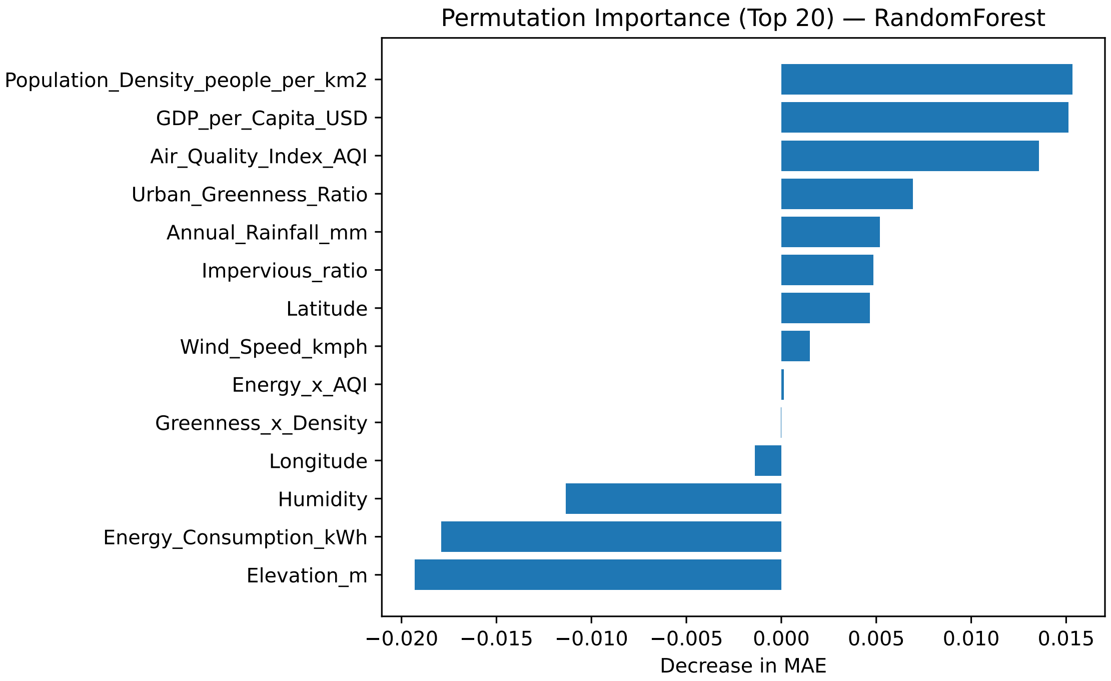

Figure 1: Permutation importance for the selected best model (Top 20 features). Larger bars indicate greater error increase upon feature permutation, i.e., higher importance.

expansion, higher-resolution land-cover products, or refined energy proxies). Alternatively, regularization adjustments and ensembling strategies can be explored to further reduce the generalization gap.

4.4 Partial Dependence

To probe functional form and directional effects, I visualize univariate partial dependence for several influential predictors. As shown in Figure 3, the relationship with `Population Density people per km2` is monotonic over the observed range, aligning with the intuition that densely built areas tend to retain more heat; `GDP per Capita USD` exhibits a mild trend, potentially proxying for infrastructure, AC load, or materials; `Air Quality Index AQI` shows a positive association that may reflect co-varying anthropogenic emissions and stagnant atmospheric conditions; and `Urban Greenness Ratio` demonstrates a cooling tendency as greenness increases, consistent with evapotranspirative and shading effects. These plots should be read as ceteris paribus curves within the model, conditioning on the empirical distribution of other variables rather than causal estimates.

4.5 Interpretation and Implications

Taken together, these results reinforce the multi-factor nature of urban heat: terrain and elevation modulate the background temperature field; morphological intensity captured by population density and imperviousness concentrates heat storage; greenness alleviates peak temperatures; and energy consumption and air quality indices reflect co-occurring anthropogenic forcings. The partial dependence curves provide an interpretable summary of direction and shape, while permutation importance ranks the practical leverage points for prediction. From a policy perspective, the combination of strong greenness effects and high importance of imperviousness suggests that targeted expansion of vegetation and reflective or permeable surfaces in dense districts should be prioritized, with site selection guided by local topography and the spatial

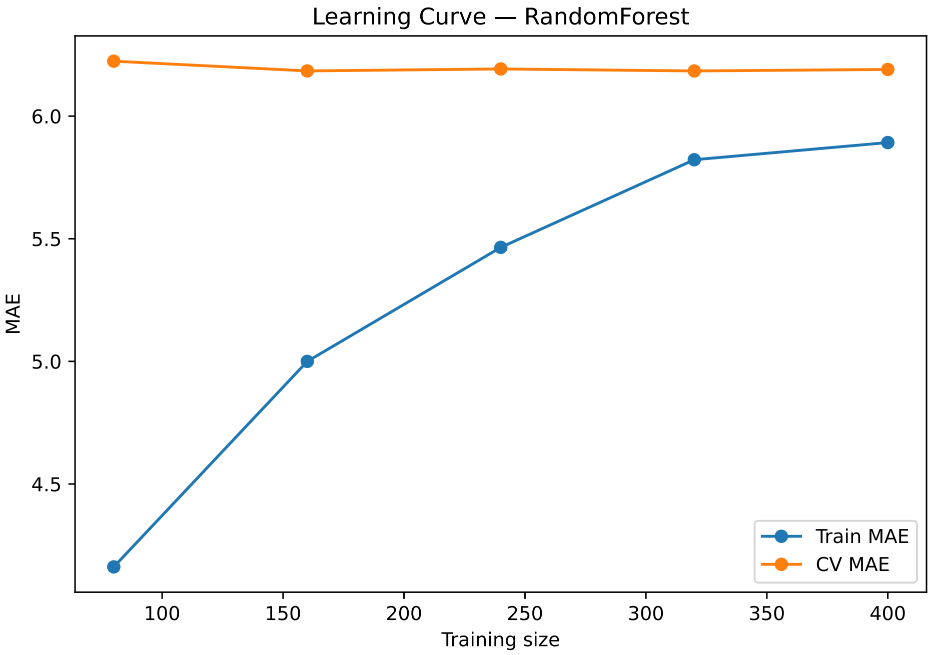

Figure 2: Learning curve of the selected best model showing train and cross-validated MAE versus training size. Convergence patterns suggest additional data could yield further improvements.

distribution of residuals.

5 Conclusion

This project presents a compatible UHI analytics stack that integrates domain-aware EDA (including elevation, slope, and aspect effects) with a principled modeling workflow. The design emphasizes transparent diagnostics, clean feature engineering and preprocessing that preserve column semantics, and repeatable model selection with exportable artifacts. This foundation supports downstream tasks such as scenario simulation and optimization (e.g., budgeted green infrastructure placement) by providing reliable predictions and interpretable drivers.

The performance across models underscores limits of thecurrent inputs, pointing to gains from enhanced features and geo-aware evaluation rather than added model complexity.

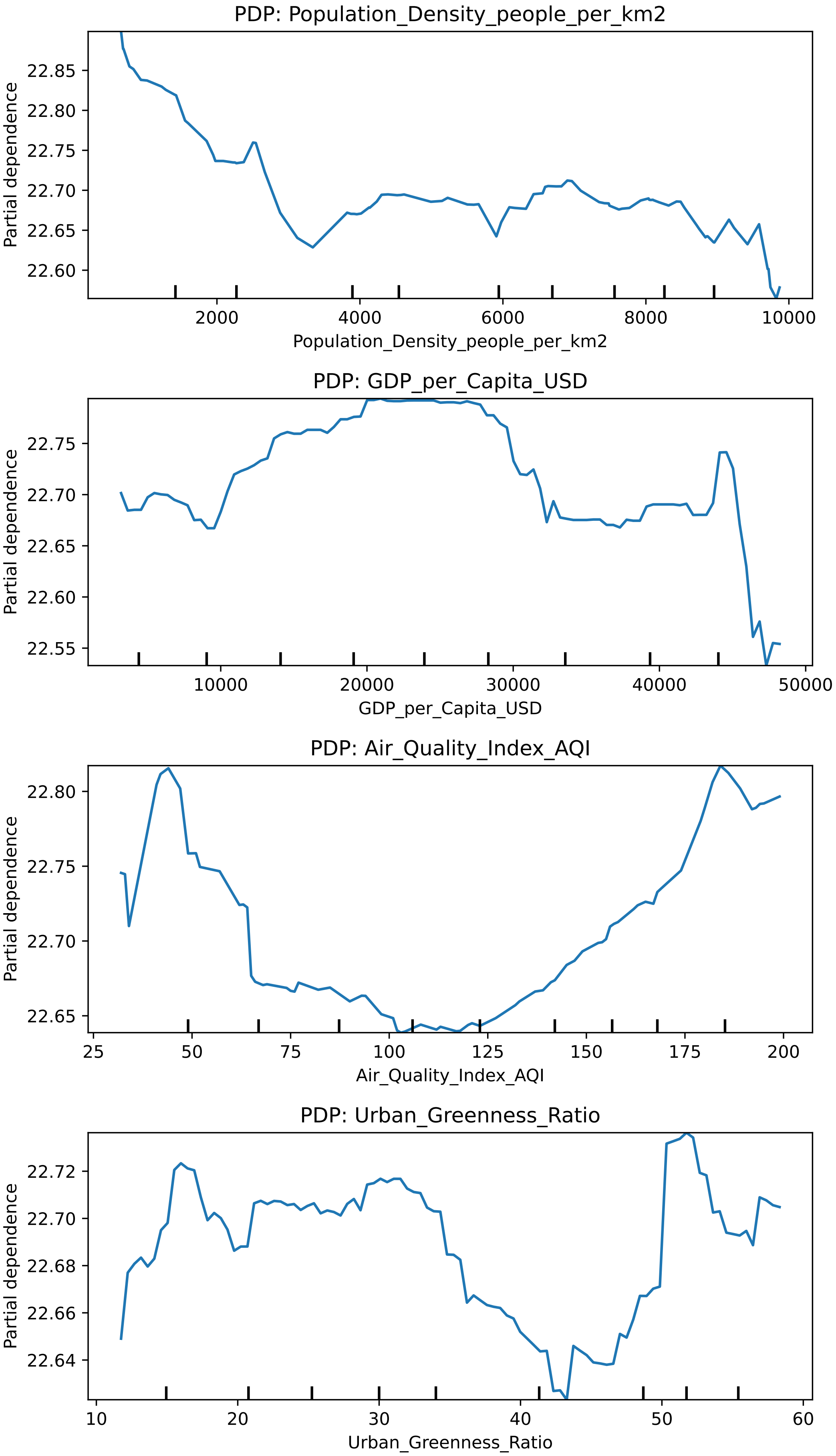

Figure 3: Univariate partial dependence for representative high-importance features. Curves indicate the model's average prediction as one feature varies while others follow their empirical distribution.