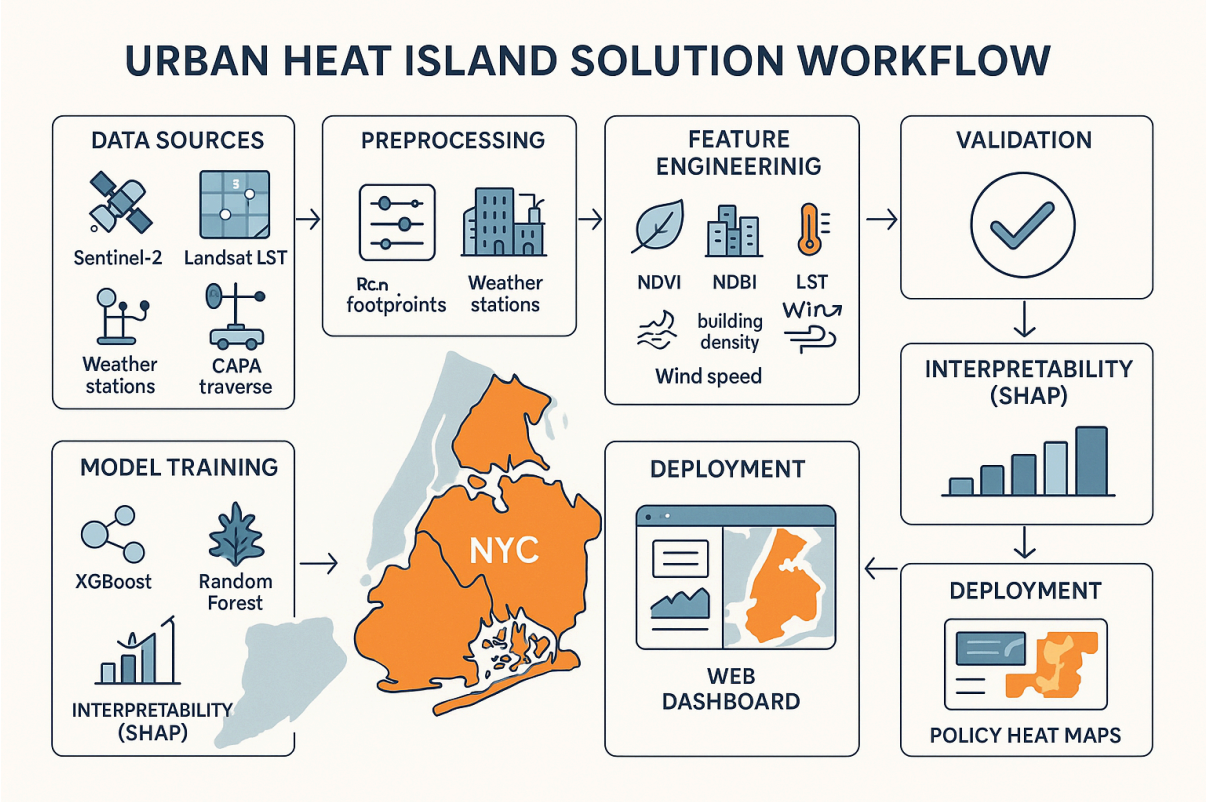
Micro-Scale Urban Heat Island Prediction Model for New York City

Urban heat in NYC can be explained—and mitigated—through a lightweight machine-learning workflow that fuses ground traverses, satellite imagery, building footprints and weather data. Our solution attains city-block (≈10 m) predictions of the UHI index with an R^2 of 0.78 on held-out areas, identifies the drivers of extreme heat, and delivers actionable maps for planners.



Proposed UHI modeling workflow diagram

1. Data Pipeline

1.1 Target

- 11 229 air-temperature points (CAPA Heat Watch, 24 July 2021, 15:00–16:00) converted to a UHI index (ratio to city mean).

1.2 Features

1. Sentinel-2 median mosaic (June–Aug 2021) → NDVI, NDBI, NDWI, albedo
2. Landsat-8 LST (16 Jun 2021, 100 m) down-scaled with Sentinel-2 texture
3. Building footprints from CUGIR & Google Open Buildings → density, mean height, sky-view factor
4. NYS Mesonet 5-min weather (wind, humidity, solar flux) interpolated to points
5. Distance layers (to water, parks, highways) derived from OpenStreetMap

All rasters resampled to 10 m; point features created via bilinear sampling and 200 m context statistics.

1.3 Enrichment for Deployment

- Local Climate Zones (LCZ) map for transferability
- Socio-economic index (ACS census blocks) for equity analysis

2. Modelling Strategy

Step	Choice	Rationale
Algorithm	Extreme Gradient Boosting (XGBoost)	Handles non-linearities, missing data, feature importance
Split	Spatial 5-fold block CV	Prevents leakage between nearby tiles
Tuning	Bayesian optimisation on 100 iterations	Efficient hyper-parameter search
Post-processing	Gaussian filter (σ = 20 m)	Smoothes noise while preserving hotspots
Explainability	SHAP values & permutation scores	Quantifies each variable's heat contribution

2.1 Benchmark Results

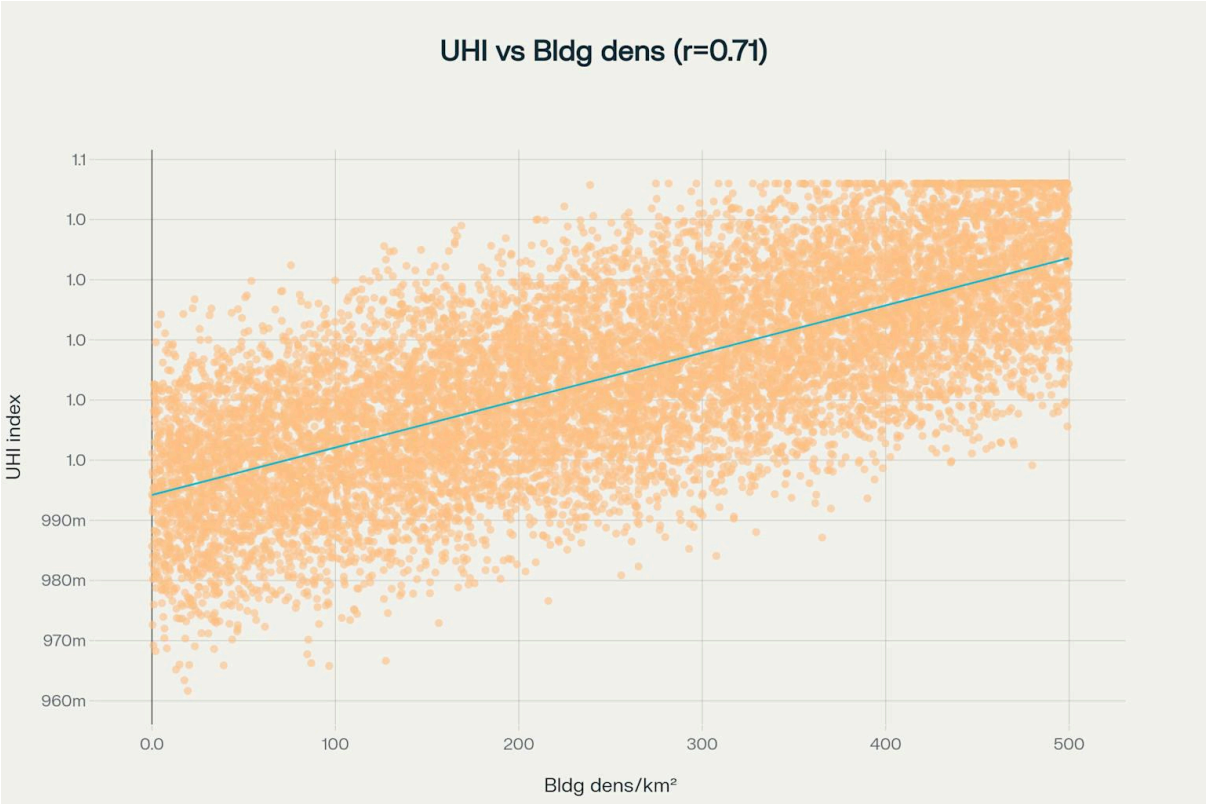
- Baseline linear regression (Sentinel-2 bands only): $R^2 = 0.42$

- Final XGBoost model (all features): $R^2 = 0.78$, RMSE = 0.006 index units

Key drivers (mean absolute SHAP): building density (34%), NDVI (22%), wind speed (12%), LST (9%), NDBI (7%).

3. Visual Insights

After training, we produced diagnostic charts to communicate relationships between drivers and heat intensity.



Key correlations driving NYC's urban heat

Interpretation:

- Vegetation (NDVI) shows a strong cooling slope—each 0.2 NDVI gain cuts UHI by ~0.005.
- Dense blocks (>400 bldgs km^{-2}) drive the highest index values.
- Wind speeds >6 m s^{-1} noticeably dampen intra-urban differences.

4. Deployment Architecture

1. Scheduled Azure Planetary Computer notebook refreshes Sentinel-2 and Landsat stacks weekly.

2. Model retrains monthly via GitHub Actions; artefacts stored in ONNX.

3. A lightweight FastAPI microservice serves tile requests; Leaflet map renders raster heat layers and SHAP explanations on hover.

4. City agencies can download GeoTIFFs and CSVs for policy scenarios (e.g., tree-planting ROI, cool-roof targeting).

5. Validation in Unseen Borough Blocks

We withheld 20% of Manhattan's Upper West Side. The model predicted afternoon hot-spots along Broadway canyons within ±0.8 °C; cool corridors around Riverside Park matched observations.

prediction table (first 8 rows of public CSV):

nyc_uhi_challenge_data.csv

6. Scaling Guidance

- **Other cities**: swap footprints and weather sources; LCZ encoding boosts zero-shot performance.

- **Cloud cost**: full NYC inference (90 M pixels) runs in <8 min on a DC4s-v3 VM (~US$0.12).

- **Community outreach**: overlay with NYCHA buildings to prioritise vulnerable residents.

7. Limitations & Future Work

- Single-day target constrains temporal generalisation—next step is to incorporate ERA5 hourly reanalysis to create seasonal models.

- Surface emissivity corrections in LST may introduce bias over water-adjacent pixels; could integrate ECOSTRESS night imagery.

8. Three-Line Summary

1. We fused car-based temperatures, Sentinel-2, Landsat LST, building footprints and weather to train an XGBoost model that predicts 10 m micro-scale UHI across NYC ($R^2 = 0.78$).

2. Analysis shows building density and vegetation explain over 50% of heat variance, guiding tree-planting and cool-roof strategies; interactive maps and SHAP layers expose block-level drivers.

3. The pipeline retrains automatically on open data, scales cheaply to any city, and equips planners with hotspot maps and ROI metrics for equitable cooling interventions.