Data:

A wide range of satellite, and urban datasets were leveraged to model the Urban Heat Island (UHI) effect in New York City's Bronx and Manhattan. These datasets capture critical drivers of urban heat, including infrastructure, vegetation, traffic, and microclimate dynamics. You can also check the github repo for this project. Below is a summary of the datasets, methodologies, and derived features:

Dataset Name	Approach	Features Extracted	URL
Landsat Satellite Data	Scaled thermal bands (LWIR11) to Celsius; weighted mosaic for temporal generalization	Surface temperature (LWIR11), thermal variation metrics, temporal stability indices.	Microsoft Planetary Computer
Sentinel-2 Satellite Data	Resampled to 30m (bilinear) to align with Landsat native res 30m (great CV & LB improvement); derived vegetation indices and surface reflectivity with weighted mosaic for temporal generalization.	EVI, Albedo, vegetation stress indices, spectral anomaly scores.	Microsoft Planetary Computer
NYC Tree Census (2015)	Aggregated tree metrics with normal and annular buffers (500m–15,000m)	Tree count, avg_tree_health, tree_size, ring features, agg features (min, max, mean, std)	NYC Open Data
NYC Building Footprints (KML/SHP)	Spatial joins across normal and annular buffers (500m–15,000m)	Building count/height/area, ground elevation (present in SHP file only), ring features, agg features (min, max, mean, std)	Shp File kml file comp dataset
NYC Traffic 2022	Aggregated traffic volume/count across both normal and annular buffers ranging from 500m–15,000m.	Traffic volume, traffic count, ring features, agg features (min, max, mean, std)	NYC Open Data
OSMNX Street Networks	Network analysis and features derived using OSMnx library.	Distance to water/park, street length, road/node density, (pedestrian, regional and vehicular areas), area types.	OSMNX Docs
NYC Planimetric Database: Elevation Points	Slope/elevation analysis with both normal and annular buffers (500m–15,000m)	Elevation, elev range, slope, agg features (min, max, mean, std)	Elevation_nyc
Weather Data	Direct sensor measurements (local weather stations) provided by competition host.	Air temperature, humidity, wind speed, solar flux, WindDir	NY Mesonet
Forestry Data	Forestry data used for inspection of stumps, tree diameters across normal and ring buffers ranging from 500m-15,000m.	Mean tree DBH, forestry count, risk score, ring features, species details, agg features	Center for Open Science

Additional Analytics: Air Pollution & Heat Vulnerability

Air pollution indices from rasters dataset and heat vulnerability index scores from csv data were analyzed during exploratory phases to assess correlations with UHI intensity. However, these were excluded from the final model evaluation, as existing features (e.g., thermal bands, vegetation indices, traffic density, urban features) already captured maximum predictive signals. This streamlined approach ensured model parsimony while retaining interpretability.

EDA & Data Cleaning

Null values, prevalent in buffer-based features (e.g., areas where buffers did not intersect buildings or trees), were imputed with 0 to reflect the absence of structures or vegetation. Radial buffers (500m–15,000m) were prioritized to encode spatial relationships without coordinates. And coordinates (Latitude, Longitude, geometry, datetime) were dropeed adhering to competition rules and regulations and for making a better generalizable model.

Feature Engineering

Many new artificial features were experimented to increase the cv score and here are some of the features that captured the mose important nonlinear patterns from the already existing features (building height dispersions & ratio features, elevation gradients, thermal mass, vertical densities, greenery building interaction and many more).

Feature Selection

Recursive Feature Elimination cv was used for feature selection, as around 1700+ features were gathered from all the above-mentioned datasets, with smaller step sizes to increase the cv score and get the most out of features, this appraoch got the most optimal features for both single model and ensembling approach.

- **Single model**: 60 features (High impact) --- (e.g., LWIR11, EVI, tree canopy density).
- **Ensemble**: 220–221 features (Dataset A/B). Smaller step sizes during RFECV maximized cross-validation (CV) gains.

Model Development Process

Many models were cross validated across same folds like randomforests, mlp, tabnet, gradient boosting regressor, histgradient boosting regressor, knn regressor, Xtreme Gradient Boosting regressor, light gradient boosting regressor, catboost regressor , and extra trees except extra trees with fine tuned parameters using optuna the above mentioned models could ony achieve cv & lb(0.96-0.9795) while Extra trees was ahead with the same folds and data with a score of 0.9824 as a single model only closely followed by xgboost with a score of 0.9791.

Hyperparameter Tuning was performed with optuna with trials 100-300 for xgboost, extra trees, lightgbm to find the most optimal parameters with a robust cross validation strategy. Parameters used in single model notebook for etr:

ExtraTreesRegressor(n_estimators=200, max_depth=55)

Single Model Strategy

The pipeline began with 1,700+ features extracted from satellite, urban, and geospatial datasets. Recursive Feature Elimination with CV (RFECV) pruned these to 60 high-impact features for single models (e.g. Landsat's LWIR11, Sentinel-2's EVI, tree cencus, forestry, urban structure). The single-model approach used an Extra Trees Regressor (ETR) tuned via Optuna for hyperparameters like n_estimators and max_depth, achieving a 10_fold CV/LB score of 0.9824.

Ensemble Model Strategy

A **10-fold Out-of-Fold (OOF)** validation framework was implemented to replicate the test data distribution and minimize overfitting to ensure that this model's ability to generalize on unseen data. The training data was partitioned into 10 folds, preserving spatial and temporal patterns. Predictions from each fold were stacked horizontally to simulate the test set's structure for predictions on only unseen data during training with only the training data. The ensemble used **27 Extra Trees Regressors** exclusively due to other models' performance gaps. These ETRs were diversified via varied parameters (depth, splits, estimators, max_features, criterion, bootstrap) to capture distinct patterns.

Initial ensemble weighting experiments tested **Hill Climbing**, **Genetic Algorithms**, and **Simulated Annealing** to optimize model contributions. These methods were ultimately abandoned due to computational inefficiency or sensitivity to initial conditions. A **Ridge regression model with a polynomial kernel (degree=2)** emerged as the optimal meta-learner, outperforming alternatives like ElasticNet, Lasso, and MLPs in stability and leaderboard performance.

- **24 ETRs** trained on unscaled Dataset A (220 features selected via RFECV)
- **3 ETRs** trained on Standard-Scaled Dataset B (221 features, including Albedo)

Before stacking, all OOF predictions (for both the training and test sets) were standardized with a StandardScaler. The scaled OOF outputs then served as input to a Ridge-kernel meta-model for final blending, ensuring a generalizable solution rather than overfitting to the test data.

Best params with ridge kernel: **KernelRidge (alpha=0.00001, kernel='poly', degree=2, coef0=0.25,)**

Final Result

This approach achieved a **leaderboard score of 0.9837**, securing 3rd place, aligned with increased cross-validation results to confirm model robustness and generalizability.

The methodology prioritized generalizability and rigorous validation to address the Urban Heat Island (UHI) prediction.