Final Report for Machine Learning

Urban Heat Island (UHI) Prediction in NYC

Yunqing Sun & Yanjun Zhou & Yifei Li

BaX 452

2025.03.10

Executive Summary:

Urban Heat Island (UHI) effects, characterized by elevated temperatures in urban areas compared to rural surroundings, pose significant challenges to public health, energy consumption, and urban infrastructure. This project aims to address the following key questions: What is the spatial distribution of predicted UHI Index values across the Bronx and Manhattan, and are there identifiable hotspots or cool zones? Which satellite-derived features are most strongly correlated with high or low UHI Index predictions? What is the potential impact of UHI hotspots on energy consumption, public health, and urban infrastructure in the identified high-risk areas? By answering these questions, the project aims to provide actionable insights for mitigating UHI effects and improving urban resilience.

In order to do this, the project is designed to predict UHI hotspots in New York City, specifically in the Bronx and Manhattan regions, using machine learning models and satellite-derived features. The dataset includes near-surface air temperature data collected on 24th July 2021, with 11,229 data points containing latitude, longitude, time, and UHI Index values. The UHI Index represents the relative temperature difference at a specific location compared to the city's average temperature, serving as the target variable for the regression model.

The project employed Random Forest, XGBoost, and LightGBM models to predict UHI intensity, with XGBoost achieving the best performance ($R^2 = 0.3136$, RMSE = 0.0141). Key findings include the identification of UHI hotspots and cool zones, with B01 (Coastal Aerosol), B12 (Shortwave Infrared), and NDVI (Normalized Difference Vegetation Index) being the most influential features. The results highlight the importance of built-up areas and vegetation in driving UHI intensity. Recommendations include implementing urban planning strategies, increasing green spaces, and using satellite data for continuous monitoring. This project contributes to the United Nations Sustainable Development Goals (SDGs), particularly Sustainable Cities and Communities (Goal 11) and Climate Action (Goal 13).

Report:

The Urban Heat Island (UHI) effect appears when urban areas experience higher temperatures than their rural surroundings due to the high density of buildings, lack of green spaces, and limited water bodies. Temperature variations between urban and rural areas can exceed 10 degrees, leading to significant health, social, and energy-related issues. Vulnerable populations, including young children, older adults, outdoor workers, and low-income communities, are disproportionately affected by heat-related illnesses. Therefore, in order to better help the society as a whole, this project aims to address the UHI effect by developing a machine learning model to predict UHI hotspots in NYC and identify key factors contributing to UHI intensity.

The dataset used in this project includes near-surface air temperature data collected on 24th July 2021 between 3 p.m. to 4 p.m. in Bronx and Manhattan, with 11,229 data points containing latitude, longitude, time, and UHI Index values. The UHI Index, which will be the target variable in this project, is the relative temperature difference at a specific location compared to the city's average temperature. Satellite-derived features, such as B01 (Coastal Aerosol), B06 (Red Edge), and NDVI (Normalized Difference Vegetation Index), were extracted from Sentinel-2 and Landsat imagery. These features capture critical environmental characteristics, such as vegetation health and aerosol presence, which are factors that could lead to UHI dynamics.

Before starting building models for analysis, EDA is provided to have a better understanding of the data structure and help gaining some key insights. Firstly, the Normalized Difference Water Index (NDWI) values were primarily negative, ranging from -0.6 to 0, indicating that the study area is dominated by urban buildings, dry soil, and vegetation, with water bodies playing a minimal role in cooling (See *Figure 1*). This explains why NDWI showed a positive correlation with UHI (0.36) instead of the expected negative correlation. Besides, the Coastal Aerosol band (B01) had the highest correlation with UHI (0.37), likely representing building reflections and urban surfaces. Similarly, the Shortwave Infrared band

(B12) and Red Light band (B04) showed strong correlations (0.32 and 0.32, respectively), indicating their importance in predicting UHI intensity.

Besides, highly correlated variables are also determined so that removal can be applied to redundant ones (See *Figure 2*). For example, the Normalized Difference Vegetation Index (NDVI) and Soil-Adjusted Vegetation Index (SAVI) were highly correlated (0.99), so SAVI was removed to avoid redundancy. Similarly, NDWI and Modified Normalized Difference Water Index (MNDWI) were highly correlated (0.84), and as MNDWI are designed for urban areas, it was retained. By having this logic, the following features were retained for model development eventually: B01 (Coastal Aerosol), which represents building reflections and urban surfaces; B12 (Shortwave Infrared), which indicates surface dryness and built-up areas; B04 (Red Light Band), which influences NDVI and vegetation health; NDBI (Normalized Difference Built-up Index), which measures urban built-up areas; B02 (Blue Light Band), which reflects urban surface characteristics; NDVI (Normalized Difference Vegetation Index), which represents vegetation cover and mitigates UHI; B05 (Red Edge Band), which influences vegetation health; B8A (Near-Infrared Band), which is used for NDVI calculations; and B07 (Red Edge Band), which affects vegetation health.

Furthermore, the spatial distribution of UHI Index values revealed distinct patterns across the study area (See *Figure 3*). High UHI Zones, represented by UHI values between 1.025 and 1.05, were typically found in dense urban areas with high heat retention due to dense buildings. Moderate UHI Zones, with UHI values around 1, were often located in urban edges or low-density building areas. Cool Zones, characterized by UHI values between 0.925 and 0.975, were associated with vegetation, water bodies, or open green spaces that mitigate UHI effects. These spatial patterns highlight the importance of urban planning and green infrastructure in reducing UHI intensity.

To further analyze the UHI and surface properties, it is important to consider the challenges associated with working with satellite data. The dataset consists of multiple Sentinel-2 spectral bands and derived indices, including NDVI (Normalized Difference Vegetation Index), NDBI (Normalized Difference Built-up Index), and NDWI (Normalized Difference Water Index). However, working with satellite data presents several challenges. First, the dataset is high-dimensional, meaning not all features contribute equally to the prediction of UHI. Second, the relationship between UHI and surface properties is nonlinear, making traditional regression methods insufficient. Third, many of the features exhibit high correlation, which introduces redundancy and potential overfitting.

To address these challenges, three machine learning models were selected. Random Forest served as a baseline model, known for its ability to handle nonlinear relationships while performing implicit feature selection. XGBoost, a boosting algorithm, refines predictions by iteratively reducing errors, improving overall model accuracy. LightGBM, designed for efficiency, allows faster training and better scalability for large datasets. Each model was trained and evaluated based on R^2, MAE, MSE, and RMSE to assess predictive performance. Feature importance was also analyzed to determine which spectral bands and indices had the strongest influence on UHI.

Among the three models, XGBoost achieved the highest R^2 value of 0.3136, indicating that 31.36% of the variance in UHI is explained. While this suggests a moderate predictive ability, it is important to note that UHI is influenced by many external factors beyond what satellite imagery captures, including weather patterns, land use, and human activities. The model's Mean Absolute Error (MAE) was 0.0111, indicating that, on average, predictions deviated by 0.0111 UHI units from actual values. The Root Mean Squared Error (RMSE) was 0.0141, further supporting that the predictions were reasonably close to actual observations.

Examining feature importance provided insight into which variables played the most significant role in UHI prediction. B01 emerged as the strongest predictor with a value of feature importance at 0.289, likely reflecting urban surface properties and atmospheric conditions (See *Figure 4*). B04, associated with heat-absorbing materials such as roads and rooftops, also had a significant influence. B8A, capturing vegetation health and the reflectivity of urban materials, and B12, linked to surface dryness and heat-retaining structures, further reinforced the role of built-up areas in driving UHI effects. Interestingly, NDVI, while relevant, was not the dominant predictor, suggesting that in dense urban environments, vegetation alone is not enough to counteract the heat retention from artificial surfaces. Similarly, NDBI, though useful in distinguishing built-up areas from vegetation, did not directly quantify how much buildings contributed to UHI intensity.

Our feature importance analysis aligns closely with our exploratory data analysis (EDA), reinforcing our initial expectations. The areas with high B01 and low NDVI and NDBI values correspond to zones dominated by heat-retaining urban surfaces, rather than significant vegetation loss. This is consistent with the background of New York City (NYC), where UHI is largely driven by dense infrastructure, reflective materials, and minimal surface moisture rather than vegetation coverage alone. The validation of our findings through both EDA and machine learning models strengthens the reliability of our conclusions.

Building on these insights, to effectively mitigate the challenges posed by the UHI effect, several strategic interventions can also be implemented. First, adopting cooler building materials is crucial. Traditional construction materials like concrete and asphalt retain significant amounts of heat, intensifying the UHI effect. By utilizing reflective or light-colored surfaces, such as cool roofs or permeable pavements, cities can minimize heat absorption and help maintain lower temperatures in densely built environments.

Second, reducing air pollution plays a key role in mitigating urban heat retention. Pollutants in the atmosphere absorb and trap heat, further amplifying temperature increases. Enforcing stricter emissions

regulations for vehicles and industries, alongside promoting cleaner energy sources, can significantly lower pollution levels and, in turn, alleviate the severity of UHI effects.

Moreover, urban planning policies should prioritize the expansion of green spaces—such as parks, rooftop gardens, and tree-lined streets—while incorporating water bodies like ponds and fountains, particularly in areas with high UHI intensity. These interventions not only provide natural cooling but also enhance overall urban livability, biodiversity, and resilience to extreme temperatures.

Looking ahead, future research should broaden the scope of this study to encompass additional cities and extended timeframes, ensuring the generalizability of findings. Integrating more environmental variables, such as humidity and wind speed, can refine predictive models and improve accuracy. Additionally, incorporating urban mobility data, such as traffic density and transportation patterns, can offer deeper insights into how human movement influences heat exposure and exacerbates UHI effects.

Finally, addressing the UHI effect requires a collaborative, multidisciplinary approach. Strong partnerships between researchers, policymakers, urban planners, and local communities will be essential in crafting sustainable, data-driven solutions that mitigate UHI impacts on energy consumption, public health, and urban infrastructure, fostering more resilient and climate-adaptive cities.

In summary, this project successfully predicted UHI hotspots in NYC using machine learning models and satellite-derived features. XGBoost performed best with an R^2 of 0.3136, and feature importance aligned with our EDA findings, confirming that urban heat retention is primarily driven by built-up surfaces rather than vegetation loss. B01, B04, B8A, and B12 were the most influential features, reflecting urban materials, surface reflectivity, and reduced moisture. Effective mitigation requires a combination of cooler building materials, reduced air pollution, and expanded green spaces to minimize heat retention and enhance urban resilience.

Appendix:

Figure 1: NDWI Value Distribution

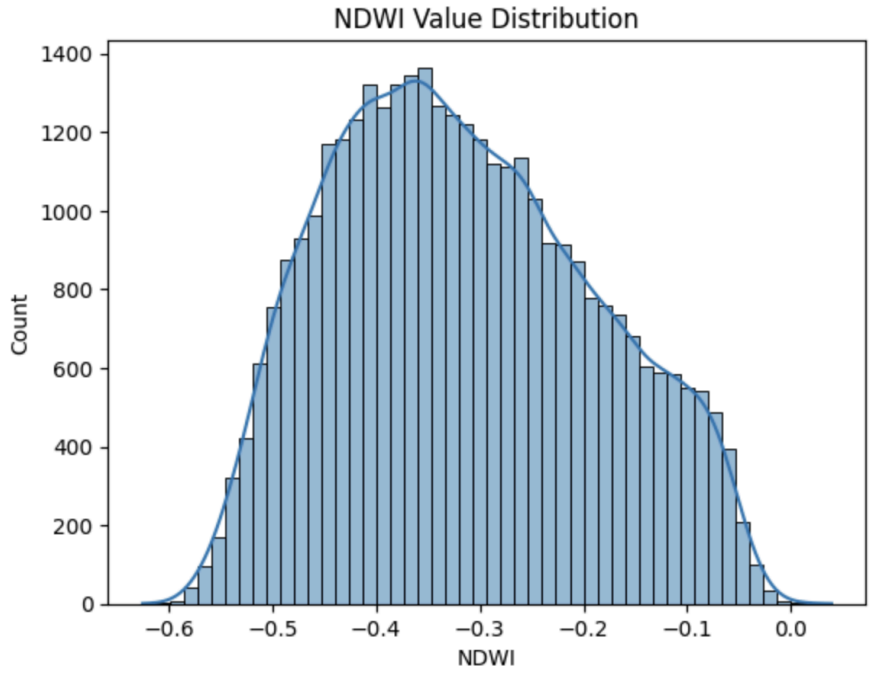

Figure 2: Correlation Matrix between Features

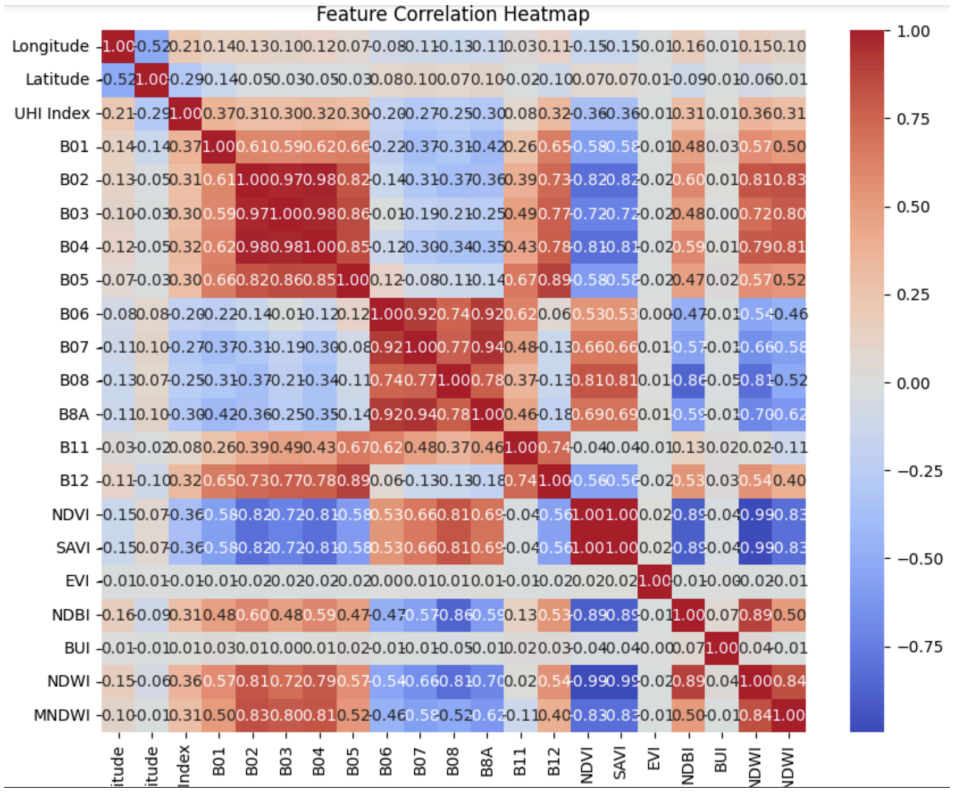

Figure 3: UHI Index Spatial Distribution

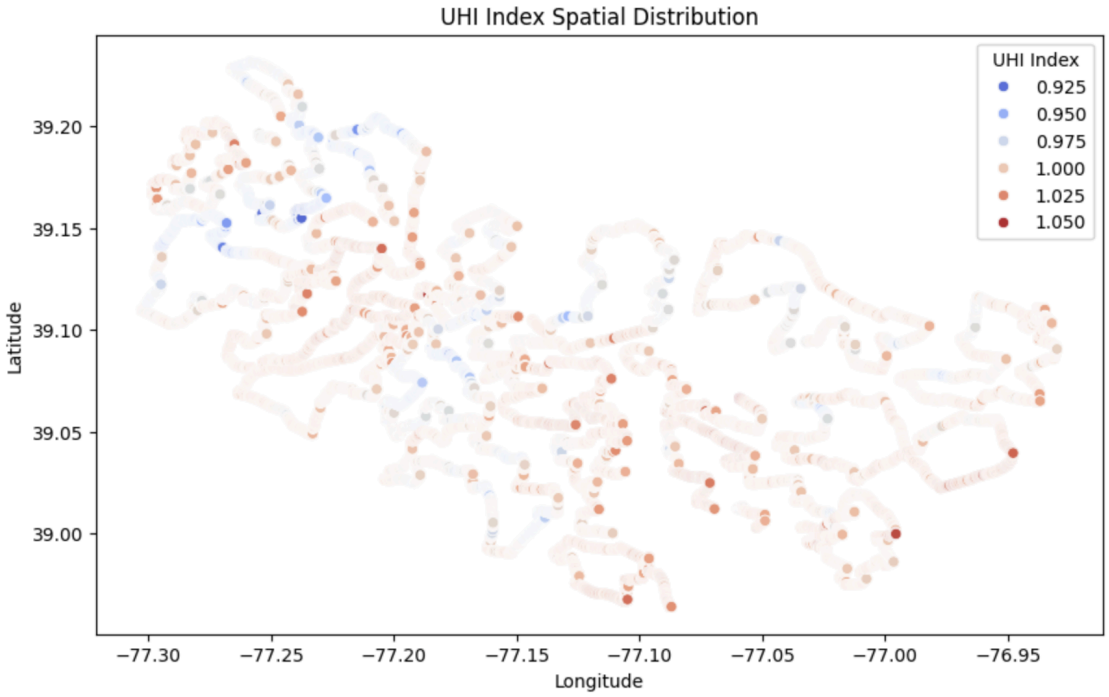

Figure 4: XGBoost Feature Importance