Machine Learning for Contributors of Urban Heat Islands in New York City

Haitian Lu

May 12, 2024

Abstract

With increasing amount of people migrating to cities, increasing development in urban areas, and increasing amount of energy used by urban infrastructures, urban heat island is not just a serious problem in big cities like New York, but has already become a world-wide problem that everyone should care about. With many contributing factors to the phenomenon, this research is aimed to guide further city planning in coordination with climate change, especially architectural planning. In order to achieve this goal, our team gather and research data from New York City urban regions regarding surface temperature, street tree distribution, greenhouse gas emissions, and building energy from 2020 to 2022. Based on the data collected, three existing machine-learning models—XGBoost, Decision Tree Regression, and Support Vector Regression— and one self-created machine learning model —Mixed Neural Network Model— were constructed and ran for further analysis. In general, the result indicates tree distribution and GHG emission respectively does not have standalone cause effects to urban heat islands, but the energy released by buildings do have strong correlation with the urban heat islands.

1 Introduction

Urban heat island is a phenomenon in which the temperature is higher in cities than countryside and the gap in the temperature between the daytime and night is larger in cities. The phenomenon is more vivid during winter and summer. Many factors contribute to the urban heat islands, and several important factors are solar radiation absorbed by the dark surface in the cities, high thermal bulk and surface radiative properties of cities' construction materials, greenhouse energy released by urban commute, lack of wind flow blocked by dense tall buildings, and the lack of green coverage. Due to the urban heat islands, the average rainfall rate in cities increases by 48 to 116 percent, the climate change fluctuates 2 to 4 times, and it causes some human health potential issues such as heat illness. An increasing number of people are concerned about the issues brought about by urban heat islands, and many researchers have already applied machine learning to look for key contributors to such a phenomenon. Researchers from South Korea found out that the green area, road-area, cropland-area, and bare area ratios were the most important factors influencing urban heat island hours, while temperature itself does not have a significant cause effect on the phenomenon (Oh et al., 2020). Since 1970, New York State's annual temperature has increased by 3 Celsius degrees, which makes extreme heat and unequal access to cool green spaces become a more serious issue than ever before and inspires out team to research green space and energy emission-related factors of urban heat islands. Three key factors evaluated by machine learning in this report are tree distribution, building GHG emissions, and building energy release. Seeing the success of using XGBoost to estimate building energy release in the paper Applied Energy (Robinson et al., 2017), I also try to run XGBoost in this research but also apply several other models and create a model by ourselves. By analyzing the result of this research, government officers can know how to adjust urban architectural planning to environmental achievements.

2 Data Collection

2.1 Surface Temperature Data

It is a cleaned data provided by the New York City Council in NYC Heat Map. It contains the mean daily temperature from June to September in 2020 to 2022. The data is originally from the satellite data from Landsat 8 produced by USGS and disseminated by Google Earth Engine. It is fine-grain data, which has 30m wide and long pixels. 27 pixels is considered as a location, and the average surface temperature is observed in every location, which means the dataset contains precise data every half mile wide and long of NYC. By using the Google Earth Engine, the water area of NYC is removed. Images with clouds from the satellite are also removed from the dataset, and the value is replaced by the average temperature of the same locations.

2.2 Tree Data

It is a data from the 2015 Street Tree Census, a census organized by NYC Parks & Recreation and partner organizations in 2015 and published in NYC OpenData. The data contains plenty of variables, including tree species, stump diameters, health conditions, and longitudes and latitudes which are most important to this research. The data covers all the street trees in each borough in NYC, which makes it a good fit for the research purpose. Among all the existing data, there are only three years of tree data, 2015, 2005, and 1995, and I chose the 2015 dataset to study our topic, considering that the change in trees will not be significant over five years.

2.3 Building Emission Data

This is a list of New York City municipal buildings over $10,000$ square feet provided by NYC Municipal Building Energy Benchmarking Results (DCAS) and published in NYC OpenData. The dataset identifies each building's address, energy intensity, and total greenhouse gas (GHG) emissions from 2010 to 2013 and is updated in May 2022.

2.4 Building Energy Data

It is energy data from Energy and Water Data Disclosure for Local Law. For the requirements of New York City Benchmarking Law, Local Law 84 (LL84), buildings that are over $50,000$ square feet need to report their annual energy and water consumption in the database. Address, total GHG, and other relevant energy release data included in the dataset are very useful to the research, and I use the dataset form 2020, which includes 3388 rows of data.

3 Method

3.1 Data Preprocessing

3.1.1 Spatial Join and Aggregation

Contributors' data (tree, building emissions, and building energy release) and aggregated surface temperature data are converted into geospatial data frames grouped by longitude and latitude, transforming these into GeoDataFrames with a coordinate reference system (CRS) suitable for GPS coordinates. The spatial join method is performed to assign each contributor to the nearest temperature reading location categorized in the surface temperature data. By counting the number of trees or amount of building GHG emission or energy release in each temperature reading location, I can analyze the correlation between these factors and the surface temperature and run regression models later.

3.1.2 One-Hot Encoding

By using the One-Hot Encoding method in the data preprocessing, I transfer the categorical data to binary, such as transferring the "Community Board" column in the data frame to the binary columns containing 1 or 0 (True or False) values, which makes us easy to analyze the data frame later.

3.1.3 Feature Scaling

The use of MinMaxScaler in the data preprocessing normalizes the feature data to the

$$[0, 1]$$

range. Without scaling, features with large values could dominate the later machine-learning model process. Following in the equation:

$$x_i = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

- x_i is the scaled value of each feature.
- x is the original value of each feature.
- x_{min} is the minimum value of each feature.
- x_{max} is the maximum value of each feature.

3.2 Model

3.2.1 XGBoost

The correlation between each potential contributor (trees, building emission, and building energy release) and the surface temperature performances bad, and I think it is possible because of many missing data in our dataset. XGBoost can learn by itself to handle missing values, so I apply it in the data preprocessing to figure out potential relationship between actual surface temperatures and those expected by the model.

3.2.2 Decision Tree Regression

Decision Tree Regression splits the data into plenty of subsets based on the features that result in the most significant reduction of variance. Once the tree is built, the predictions are made by passing the decisions in each node about going left or right until a leaf node is reached. The reason I choose to use Decision Tree Regression in the model is that they can handle non-linear relationships.

3.2.3 Support Vector Regression

It performs high dimensional feature space using kernel and finds a flat function that has the most ϵ deviation from the actual value. Because of its high-dimensional nature, it is good at dealing with non-linear relationships, and that is one of the reasons I apply to this model. The other reason is that Support Vector Regression provides good robustness in outliers, because it can choose to fit in certain margins, so extreme values from the dataset will not have strong influence on the model performance.

3.2.4 Mixed Neural Network

It is a self-created neural network model that handles both continuous and categorical data. By processing through embedding layers, categorical features can concatenate with continuous features and establish a new input to process through the following hidden layers. The outputs that come from hidden layers are evaluated by Mean Squared Error (MSE) as the loss function, and the model will keep changing the weight matrix in the hidden layers to minimize the loss. The Figure 1 shows the structure of the Mixed Neural Network Model.

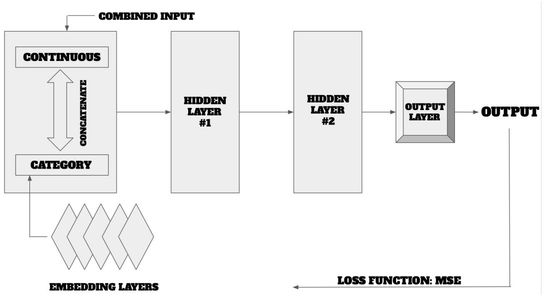

Figure 1: Structure of Mixed Neural Network Model

3.3 Cross Validation

It is a technique used to estimate the performance of a predictive model and is largely used when the dataset is not large enough. The dataset would be separated into k folders, and the model would be trained on the data from $k-1$ folders and validated by the remaining folder. By using this technique, each folder is used as the test sample once and used as a training sample for $k-1$ times. The mean or median result from the k folders provides a more accurate estimate of the model performance compared to a simple train/test split.

3.4 Bayesian optimization

It is a highly effective method for fine-tuning the hyperparameters of the Mixed Neural Network model. Because of requiring fewer evaluations of the objective function compared to traditional methods, it can efficiently handle high-dimensional spaces and sampling. Defined at the beginning of the process, a function that integrates the training and evaluation procedure uses given parameters to train the model and returns the negative value of the average test loss to minimize the test MSE. Then, the optimizer continuously selects hyperparameters sets by balancing unexplored regions and pre-randomly explored regions to predict the performances of different hyperparameter combinations.

4 Result

4.1 Tree

I locate the center of each location and convert them to a CSV file containing surface temperature data grouped by longitude and latitude. I use spatial join method to find the closet temperature reading location to each tree, which helps us find out the number of trees in each location. Because of the 0.0983 Pearson correlation coefficient calculated and the scatter plot shown in the Figure 2, I think trees are uncorrelated to the surface temperature. I also try to use XGBoost to handle potential bad performance caused by missing values, but it still performs badly with 1.96 Root Mean Square Deviation (RMSE), and the relationship between the predicted value and the actual value is shown in the Figure 3. At the end, I conclude that the trees are uncorrelated to surface temperature.

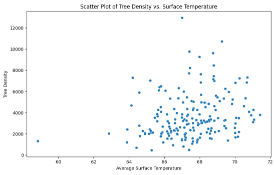

Figure 2: Scatter Plot of Trees Density Distribution in Temperature

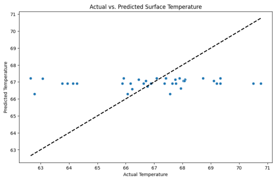

Figure 3: Predicted Temperature from Tree Data in XGBoost

4.2 Building GHG Emission

Similar to the way of testing the correlation between street tress and surface temperature, I use the spatial join method to assign buildings to their closet temperature reading locations and average the emissions of all buildings in each location. Because of the poor performance shown by the -0.058 Pearson correlation coefficient, I filled the nah values by the mean of the column and then choose to run Decision Tree Regression and Support Vector Regression. The performance of both models is shown in the Figure 4 and Figure 5, and because of their poor performance, I conclude that the building emissions are uncorrelated to surface temperature.

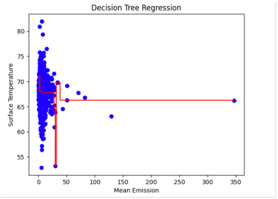

Figure 4: Decision Tree Regression: Distribution of Temperature in Emission

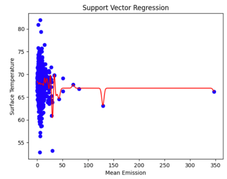

Figure 5: Support Vector Regression: Distribution of Temperature in Emission

4.3 Building Energy

By calculating the correlation of each feature and value, I choose features with absolute correlation value larger than 0.1. Through One-Hot Encode and MinMaxScaler, I convert "Community Board" to binary columns and normalize the feature data to $[0, 1]$ to ensure that all input features have equal effects on the operations by the later machine learning models. Two models are used here:

4.3.1 Decision Tree Regression

I use R^2 and MSE to evaluate the performance of Decision Tree Regresion. The model performs well at the end, with $0.4466R^2$ score and 7.0414 MSE score. The relationship between the expected and actual surface temperature value is shown in the Figure 6.

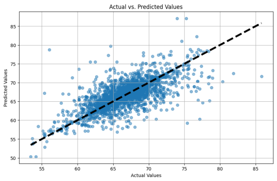

Figure 6: Predicted Temperature from Energy Data in Decision Tree Regression

4.3.2 Mixed Neural Network

Before running the Mixed Neural Network model created by ourselves, I use LabelEncoder to covert categorical labels into numbers and separate the data into continuous and categorical data. I choose to use the model with 3 embedding layers and 2 hidden layers. The results show that it has 14.127 MSE score and the relationship between the expected and actual surface temperature value in shown in the Figure 7. Then, shown in the Figure 8, I also apply Cross Validation with 5 folders to evaluate the performance of the model, and the mean of MSE scores from 5folders is clearly lower than 14, which means the model's performance is actually better than the 14.127 MSE score from simple train/test split.

In the end, I apply Bayesian optimization to find the best-fit model. I define the space of hypermeters in different learning rates, number of layers, batch sizes, and hidden sizes. By the end of the tuning process, the optimizer identifies a best-fit set of hyperparameters that yields the lowest test MSE. The set contains 0.0345 learning rate, 86.6890 embedding size, 63.8508 hidden layer 1 size, and 26.7634 hidden layer 2 size.

Based on either the MSE score or the visual relationship of the expected and actual value, the Decision Tree Regression performs much better than the model built by ourselves.

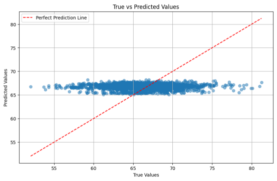

Figure 7: Predicted Temperature from Energy Data in Mixed Neural Network

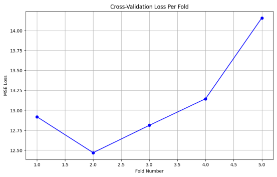

Figure 8: MSE of temperature in Cross Validation

5 Result & Discussion

This research aims to identify the correlations between urban features—specifically trees, building emissions, and building energy outputs—and urban heat island effects in New York City, employing a comprehensive framework supported by multiple machine learning models. Taking surface temperature as an indicator of the urban heat islands' effects, the research analyzes the correlation between potential features and surface temperature one by one.

As evidenced by a Pearson correlation coefficient of 0.0983 and an RMSE value of 1.96, the research suggests that tree coverage does not significantly impact surface temperatures. Building GHG emissions also show no significant correlation with surface temperatures, indicated by a Pearson correlation coefficient of -0.058. Despite attempts to upgrade the performance by using different regression models, including XGBoost, Decision Tree Regression, and Support Vector Regression, the outcomes still demonstrated minimal impact on surface temperature. Two models, Decision Tree Regression and Mixed Neural Network Model, run in the analysis of building energy suggested some influence on the surface temperature. Outperformed by the Mixed Neural Network Model, Decision Tree Regression runs a result of $0.4466R^2$ Score and 7.0414 MSE score, but the relationship is not overwhelmingly strong.

Several factors may need to be further investigated to upgrade these models and conduct better research on this topic. First, different indicators could be used to as urban heat island effects, and indicators observed twice during daytime and night are better fit. Although our research uses the average surface temperature as the indicator of the high temperature in cities compared to rural regions which is one key phenomenon of urban heat islands, but the difference of temperature between daytime and night in cities, the other key phenomenon, is not reflected. If datasets like daily air temperature of daytime and night are provided, the model will have a much better performance. Second, many other potential contributors to urban heat islands could be added, such as the density of tall buildings and the density of commute, and the correlation between the phenomenon and the combination of other effects should be analyzed. Third, factors that reduce urban heat islands should be analyzed as well. It is possible that plenty of factors that contribute good and bad to urban heat islands exist in the same location, which minimize the correlation result of certain factors. For example, it is possible that the large size of green area in Central Park reduces the urban heat island effects around middle town in Manhattan, which lowers the buildings GHG emission's influence on urban heat islands and makes researchers hard to compare it with places with less buildings and lower urban heat islands phenomenon. Fourth, if more recent data is available, the performance will be better. The building GHG emission data I use is from 2010 to 2013, which is 10 years from the date of the research, but it is the most recent data I can find. If the government can disclosure the data from 2020 to 2024, the model will performs much better, and the result will be more significant.

Overall, the study suggests that individual elements such as tree coverage and building emissions might not have a significant standalone impact on the urban heat islands and indicates the multifaceted nature of such phenomenon. Through the strong relationship between building energy and urban heat islands, government policymakers can gain more insights into managing building energy utilization.

Works Cited

2015 Street Tree Census. https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/pi5s-9p35, NYC OpenData, Department of Parks / Recreation, 2022.

Energy and Water Data Disclosure for Local Law. https://data.cityofnewyork.us/Environment/Energy-and-Water-Data-Disclosure-for-Local-Law-84-/r6ub-zhff/about_data, NYC OpenData, Department of Buildings, 2024.

NYC Heat Map. https://github.com/NewYorkCityCouncil/heat_map?tab=readme-ov-file#nyc-heat-map-surface-temperature, New York City Council, 2022.

NYC Municipal Building Energy Benchmarking Results. https://data.cityofnewyork.us/City-Government/NYC-Municipal-Building-Energy-Benchmarking-Results/vvj6-d5qx/about_data, NYC OpenData, Department of Citywide Administrative Services, 2022.

Oh, Jin Woo, et al. "Using deep-learning to forecast the magnitude and characteristics of urban heat island in Seoul Korea". *Scientific reports*, vol. 10, no. 1, 2020, pp. 3559–59.

Robinson, Caleb, et al. "Machine learning approaches for estimating commercial building energy consumption". *Applied energy*, vol. 208, 2017, pp. 889–904.