Assignment 4 – DATA 622

Overview

For this assignment I decided to leverage my capstone project to also cover the requirements for this DATA 622 Final Project. Therefore, while this file attempts to place everything needed into one coherent summary file, the Brightspace submission will have several additional links in order to provide full context if needed. The dataset used in the final analysis is one that was custom built from a variety of publicly available datasets, as well as using multiple different enrichment techniques. In short, this document is a condensed version of my capstone project to submit for DATA 622 Assignment 4.

The Problem

There has been a lot of research into the presence of urban trees on the surrounding environment, specifically on their impacts on countering extreme heat and cold. Specifically, urban trees have been found to damper the effects of extreme head through providing shade and as a result of evapotranspiration. Additionally, there have been findings that outline how the presence of trees can help block wind, which tends to damper the impacts of extreme cold as well. As weather extremes become increasingly common due to climate change, further study on the impact of urban trees is needed. Specifically, in different regions in order to better understand how trees can help different cities. For this project, data specific to New York City was used in order to help identify any potential impact trees can have on energy use intensity in buildings.

Preparing the Data

This project uses a custom built dataset derived from a multitude of different public datasets and resources. In short there were three main phases of gathering, enriching and processing data in order to distill the final working dataset.

Part 1 - Buildings Data & Energy Usage (BuildingWork_Part1.ipynb)

The first section of the data made use of Local Law 84 energy benchmarking data, which was is ingested via the NYC Open Data Socrata API for all years available, but ultimately was restricted to 2010 and 2017. This restriction was done to align with the primary second data source, which was the tree canopy change data. The main feature of interest in the Local Law 84 data was the "weather normalized eui" information, which is the energy usage intensity for each building that reports data normalized to counter weather variations between years. The raw data included a large number of administrative and reporting fields, which needed to be sifted through and dropped. The data was filtered in order to keep

observations that are metered at the whole-building or whole-property level, while allowing some metering fields to remain null, as these fields were introduced in later reporting years. The buildings that were kept for this analysis, included those that were predominantly residential properties. The identifying dimensions of buildings, like unique identifiers (e.g., BBL, property id, addresses) were enriched using various methods like self-joining and public APIs. Further more, these identifying columns, once properly enriched, were used in order to geocode each building to obtain a latitude/longitude point where the raw data had nulls. These geographic points were used in order to further add to the data with census tracts, building footprint geographies, and other spatially oriented data. Additional building information, such as Number of floors, zoning classifications, construction year, etc. were added to the data via the MapPLUTO API. Lastly, LiDAR-derived canopy change data that classifies the city into areas of canopy gain, loss, or no change between 2010 and 2017 was ingested, and subsequently joined to the building data in order to categorize each building into one of those canopy categories. A buffer space of 50 feet was used, in order to limit any shifts in canopy coverage to within 50 feet of a building's footprint geometry. The final data set from this section limited to buildings that intersected (with the 50 foot buffer considered) with the canopy data for analysis.

Part 2 - Tree Count Data (TreeWork_Part2.ipynb)

The second section of processing focused on city-level forestry and tree data. As another tree-focused dataset, NYC street tree inventory data and forestry work order data from NYC Open Data were also ingested and processed. The data was filtered to keep only trees plausibly present during the 2010–2017 period. Newly planted trees that appear only very late in the series are dropped as too young to have a meaningful effect on shading or wind, while dead trees, stumps, and records with unusable coordinates are removed entirely. Finally, forestry work orders data, which is another tree focused dataset, was used to find and flag tree removals yielding a shift in tree inventory between 2010 and 2017. The yielded data is an attempt at a tree count for NYC in 2010 and 2017, two years that are outside of that "NYC Tree Census" years.

Part 3 - Putting it Together & Aggregation (Analysis_Part3.ipynb)

The third and final steps to the data preparation and processing phase, is where the Part 1 data and the Part 2 data are put together into a final product. Essentially, using a spatial join, the tree counts were joined to the building data. Similar to the canopy data any tree that was within 50 feet of a building geometric footprint was included in the data. In short, this step allow for the aggregation of a tree count for each building within the dataset to help with impact analysis.

The last series of steps taken to yield the ultimate working data set was a series of filtering and processing steps that were as follows:

- *Removing Outlier weather_normalized_eui values.*
  Those buildings with weather_normalized_eui values that were found to be above the 99th percentile and below the 1st percentile were removed in an attempt to improve the distribution of the values.
- *Limited the data to those buildings in class C or D zoning class*
  This was to keep the building pool to a uniform multifamily residential building type in an attempt to further control for variation. Class D and C are residential apartment buildings with and without elevators, respectively.
- *Limited the data to those buildings 6 floors or below*
  Limiting to 6 stories or under was the smallest residential building size available in the Local Law 84 data.
- *Limited the data to those buildings that were built after 1900*
  Removing buildings that were built before 1900 to help control for building style and history, additionally there were only a few of these instances.
- *Feature Engineered a "commercial_floor_flag"*
  Using the building class information, those buildings that were zoned for first floor commercial use were flagged (1). This was to differential from purely residential buildings.

This file, which includes the last mile of processing along with the exploratory data analysis and the modeling section itself, has two distinct modeling sections. The first one, which leverages a series of Difference in Differences modeling to attempt to identify the impact of canopy change on building EUI was the initial modeling for the capstone project. The second series of models, which use Cross Validation with Ridge and Lasso regression methodologies, was more aimed at this DATA622's Assignment 4 and its mandates. These models, outlined below, attempt to build a model that would predict change in weather normalized EUI based on the features in the dataset.

Data Modeling and Analysis

First Series of Models (Difference in Differences OLS Models)

A series of nine difference-in-differences (DID) OLS models were put together using weather normalized site EUI and various combinations of transformations (e.g., log, square-root) and features. In each of these models the canopy change class was treated as the key variable. Across all models the r-squares values were poor and the fits were weak. None of the tree features' coefficients were statistically significant, most of the building-focused features were also not significant. However, the one consistently significant

covariate was the 1950–1980 year-built bracket. This suggests that the age of buildings influences eui. Overall, these models indicate that, within this dataset and time window, canopy change alone does not show any noticeable effect on building energy use.

Table 1 – Difference In Differences (DiD) OLS Modeling Results Overview

Model Number	Formula	R^2 Value / Adj. R^2 Value	AIC / BIC	F-Statistic / Prob. (F-Statistic)	Statistically Sig. Coefficients (p-value <0.05) & Notes
1	weather_normalized_site_eui ~ post_2017 * canopy_change_class + NumFloors + UnitsRes+ ground_elevation+ height_roof+ C(year_built_bracket) + C(commercial_floor_flag)	0.037 / 0.026	9831 / 9895	2.439 / 0.00431	C(year_built_bracket)[T.1950-1980]
2	log_eui ~ post_2017 * canopy_change_class + NumFloors + UnitsRes+ ground_elevation+ height_roof+ C(year_built_bracket) + C(commercial_floor_flag)	0.035 / 0.024	1065 / 1130	2.090 / 0.0162	C(year_built_bracket)[T.1950-1980]
3	sqrt_eui ~ ~ post_2017 * canopy_change_class + NumFloors + UnitsRes+ ground_elevation+ height_roof+ C(year_built_bracket) + C(commercial_floor_flag)	0.037 / 0.026	3944 / 4008	2.335 / 0.00645	C(year_built_bracket)[T.1950-1980]
4	weather_normalized_site_eui ~ post_2017 * canopy_change_class + NumFloors + C(year_built_bracket) + C(commercial_floor_flag)	0.032 / 0.024	9830 / 9880	2.658 / 0.00509	C(year_built_bracket)[T.1950-1980]
5	log_eui ~ post_2017 * canopy_change_class + NumFloors + C(year_built_bracket) + C(commercial_floor_flag)	0.030 / 0.022	1065 / 1114	2.122 / 0.262	C(year_built_bracket)[T.1950-1980]
6	sqrt_eui ~ post_2017 * canopy_change_class + NumFloors + C(year_built_bracket) + C(commercial_floor_flag)	0.032 / 0.023	3943 / 3993	2.449 / 0.00978	C(year_built_bracket)[T.1950-1980]
7	weather_normalized_site_eui ~ post_2017 * canopy_change_class	0.001 / -0.003	9854 / 9884	0.4113 / 0.841	N/A
8	log_eui ~ post_2017 * canopy_change_class	0.002 / -0.003	1087 / 1116	0.4514 / 0.812	N/A
9	sqrt_eui ~ post_2017 * canopy_change_class	0.001 / -0.003	3967 / 3997	0.3788 / 0.863	N/A

Second Series of Models (Ridge & Lasso Models)

There were a total of three different models created the first a simple Ordinary Least Squares (OLS) Model as a baseline comparison, and then a lasso regression model and a ridge regression model. Both of the latter two models used a cross-validation method of 5 different folds of the data. Furthermore, for the alpha values used in the Lasso and Ridge models, we generate an array of 100 logarithmically spaced values between 0.001 and 1,000 using np.logspace(-3, 3, 100). Each of these alpha values is evaluated to find the value that best balances for the bias–variance trade-off and thus is considered the best regularization strength for the model. Of these models the Ridge method with an alpha value of 572.236 yielded the best model. The Lasso model was yielded nearly the same r-squared value, but had higher error values (MSE, RMSE, and MAE). While the Ridge model was the relative best, none of these models are actually good at predicting weather normalized EUI shifts with the features available, the ridge model is essentially the least worst.

Table 2 – Ridge & Lasso Linear Modeling Results Overview

Model Number	Method	R^2 Value	Found Alpha	RMSE/ MSE / MAE
1	OLS Model	-0.2282	N/A	1230.177 / 35.074 / 24.516
2	Ridge Model (cv = 5)	-0.0305	572.236	1032.155 / 32.127 / 22.089
3	Lasso Model (cv = 5)	-0.0341	1.072	1035.806 / 32.184 / 22.087

Conclusion & Potential Impact

Good faith efforts were made at constructing models that detect and predict shifts in a building's weather normalized EUI. Neither the DiD models or the predictive models (OLS, Ridge, and Lasso) did a decent job. With the DiD models, the overall fits were low and none of the features, with the exception of the 1950–1980 year-built bracket, were statistically significant. This suggests that changes in energy use were not impacted by the changes in canopy around a building for this period. All the predictive models produced negative r-squares values with the 20% test set. This implies that truly predictive features are not present in this dataset or inter-feature relationships that are not linear. Potential future work in this area would be to collect and use more granular data, as well as hopefully having data on smaller buildings that would more likely be impacted by tree presence.

If any of these models had substantive results, the impact on urban planning and targeting specific areas for tree planning could have helped New York City reach their climate goals. Buildings are the number one emissions source in the city, and if these models demonstrated tangible impact, trees would have been a simple but significant means of

improving energy efficiency. Again, if the results were substantive, decision makers would have another tool at their disposal when drafting up zoning changes, policy proposals, etc. In short, decent models would help link urban forestry and energy policy together, strengthening the case for urban tree planting as a climate resilience strategy.