ML & Climate | Predicting NYC Street Tree Health

Kate Harwood krh2154

May 2, 2022

The importance of city greenery is growing with the increasing urbanization of the world. Urban trees provide many benefits like temperature regulation, stormwater runoff reduction, and air pollutant removal [9]. However, trees must grow to maturity before many of these advantages can be realized [10]. Keeping vegetation thriving in a city is difficult, especially so for street trees that cannot benefit from an urban park ecosystem. From living in concrete root beds in the shadows of tall buildings, to serving as pets' toilets and humans' trash cans, urban trees face many obstacles to staying alive and healthy [10]; the average life of a street tree in a dense urban environment is only 13 years [8]. Gaining information on what factors affect the health of urban street trees can help facilitate the growth of vitally important green ecosystems in our cities.

This research shows that the future health of New York City's half a million street trees can be predicted on a relatively granular level with fairly high accuracy using data collected from NYC street tree censuses and publicly available land use data. Given information about trees and the surrounding buildings in a quarter block radius, my models can predict the average health of trees in the given area 10 years later with greater than 90% accuracy.

A variety of models were trained on different feature groupings, with the highest model achieving an accuracy of 91.7% and a Macro F1 score 85.1% on a 4-way categorical health status classification task. Location information and nearby building size are shown to be the most important features for accurate prediction.

Data

The primary data used in this study are from three separate censuses of NYC street trees in 1995, 2005, and 2015 [1] [2] [4]. These censuses, organized by NYC Parks and Recreation and others, are administered every ten years by volunteers, and contain information about all of the street trees in New York City's five boroughs, (not including trees in the city parks). The data includes key features, such as latitude and longitude of trees, tree species, and perceived tree health. Other "accessory" data points include features such as presence of wires on or around the tree (trunk, stump, house tap, secondary, etc), sidewalk condition around the tree (cracked, raised), presence of debris in tree branches (shoes, lights, etc), and others. Figure 1 shows examples of these features. Additional features from the PLUTO (Primary Land Use Tax Plot Output) NYC dataset are also used [7], incorporating information about nearby buildings into the prediction task.

Feature Matching

There are many "accessory" categorical tree condition features in each ten-year dataset, however the features do not match exactly across datasets, which limits the usable features for this study. The two "accessory" categorical features that were matchable with the least amount of data distortion were sidewalk condition and wire presence, however even these features required some generalization to be usable across datasets. For instance, the 1995 census recorded whether the sidewalk was cracked or raised in one categorical variable, while the 2005 dataset recorded cracked sidewalks and raised sidewalks in two separate variables. For this study, these two variables were combined into one categorical variable to match the 1995 sidewalk variable. The wire feature had even more disparate variables tracking types and location of wires across

Figure 1: Examples of features recorded in tree census.

the two datasets, requiring more generalization to create a feature that matched across the datasets.

Tree Matching

Once the feature data were made compatible across datasets, the datasets were combined. 1995 features were combined with 2005 tree health as targets, and 2005 features were combined with 2015 tree health as targets. The resulting two datatables were concatenated together to create the final dataset.

Since the trees are not given unique ids to match across datasets, they need to be matched by location instead in order to complete this merging. The latitude and longitude of each tree is not exact, and therefore cannot be used as an exact matching mechanism.[1] Instead, the trees were binned according to latitude and longitude, and the bins were matched across datasets. The 2005 data was binned into latitude and longitude buckets with a step size of 0.0005 (about 55 meters when measured at the equator, or a quarter of a city block in NYC). Data that did not match on zip code across datasets were removed. The 1995 and 2015 data were binned into the same buckets as the 2005 data, using latitude and longitude to match to the nearest bin. Data points from the 1995 and 2015 datasets with latitudes and longitudes further away than 0.001 from the location of their bin were thrown out (111 meters at the equator, or about half a city block in NYC).[2] There were 86,858 unique lat/lon bins created for all of NYC.

Finally, the health statuses of the trees were averaged across the trees in each bin, and this became the tree health label for the given bin.

PLUTO Data

The PLUTO data was filtered for data points that included latitude, longitude, zip code, number of floors, land use category, year built, and did not include buildings constructed after 2005, since the task is to predict tree health in 2005 and 2015 using features available before those respective years. The PLUTO data was then binned into the existing latitude and longitude buckets. For the rare cases where there was more than one PLUTO data point in a bin, the number of floors and land use category were averaged across the data points, to ensure there was only one PLUTO data point per bin. This was so the PLUTO data could then be joined on latitude and longitude bin to the final tree data. Any latitude and longitude bins from the tree data that did not have corresponding PLUTO data were thrown out. This reduced the number of data points dramatically from 1,008,859 to 3,091, however better results were still achievable with the PLUTO data than without.

[1] From the tree census data guide, it seems the geographical marker collected was the address of the building nearest to the tree, which was later mapped to latitude and longitude coordinates in an imperfect process. As the data guide states of these final coordinates: the trees are "near to it, across the street from it, or around the corner from it." [3]

[2] The number of trees in each bin varies quite a bit. There are some bins that have one or a few trees, and some with over 100 trees (and one with over 300 trees), meaning there may be some data collection error when assigning latitudes and longitudes to the collected trees. In the data used in the final model, which includes the PLUTO data, there are fewer bins with a concerning number of trees in them. Still, this could be something to explore further.

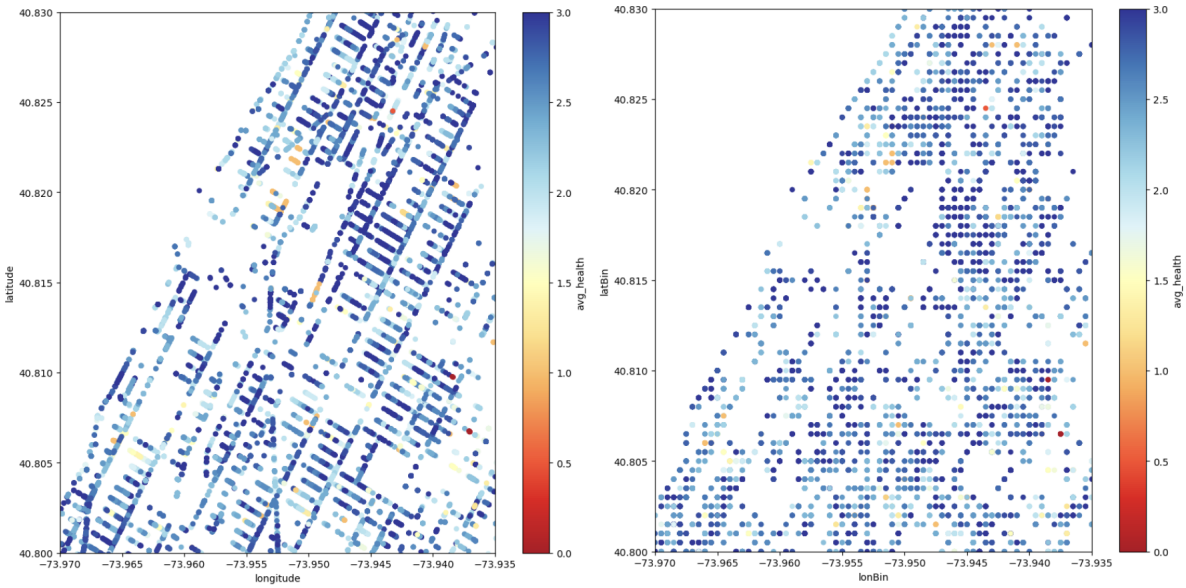

Figure 2: Section of the west side of midtown Manhattan showing the health of individual street trees (left) and the average health of the trees in each latitude and longitude bin (right). Higher on the color scale indicates better health.

Final Features

Species (36 unique values): Latin species name[3]

Diameter (40): Diameter at breast height of tree

Sidewalk cond. (2): If the sidewalk is cracked or raised around tree, or in good condition

Wire presence (2): If there are wires present above tree or interfering with tree, or no wires

Lat bin (130): The latitude bin containing the tree

Lon bin (109): The longitude bin containing the tree

Num floors (52): The number of floors in the building in the bin (avg where necessary)

Land use (11): The land use category for the land in the bin (avg where necessary)

Tree Health (4): Target feature, average health of trees in bin

Methods

Because the data is highly imbalanced (as shown in Figure 3), the smaller categories were upsampled to create a balanced training set (test data was not upsampled). This yielded better results than downsampling. The data was split into 80% training data, 20% test data. Different models and training methods were applied to the data: Linear Regression, Random Forest Regression, Logistic Regression, Random Forest Classification, Gradient Boosting Classification, and an SVM. Experiments were conducted with "identity features" (latitude bin, longitude bin, zip code, species, diameter), with "identity features" + "accessory features" (sidewalk, wires), and with "identity features" + "accessory features" + PLUTO features (land use category, number of floors). The regression models predicted the average health of trees in a bin on a continuous

[3] There are hundreds of tree species in the original tree datasets, so rather than one-hot encoding them, the species are encoded as their frequency count. This results in some loss of info, since certain species will have the same frequency count. There is a trade-off between loss of info and too many feature values.

Class	*Count*
Excellent	918
Good	1447
Fair	102
Poor/Dead	6
Total	3,091

Figure 3: Total data points with PLUTO data included.

scale from 0 to 3. The classification models predicted the average health of trees in a bin rounded to the nearest health status class number (0: Poor/Dead, 1: Fair, 2: Good, 3: Excellent).

A combination of hyperparameter grid search and manual finetuning was performed for a slight boost in scores on the most promising models (RF Classifier and GB Classifier) over default hyperparameter settings. Final hyperparameters for the Random Forest Classifier included 1,000 decision trees and used entropy as the split criterion. Final hyperparameters for the Gradient Boosting Classifier included 1,000 decision trees, a learning rate of 0.01, and a max depth of 10.

Validation was performed on the two best models by training from scratch with five different train/test splits. Average scores and standard deviations on test data were reported for these models.

Results and Discussion

Figure 4 shows the results for each model trained and tested once. The top of the table shows results from models run without the PLUTO data, and the bottom of the table shows the results from models run with the PLUTO data. The classification models performed much better than the regression models, and the RF models performed better than the Linear and Logistic models on the same data.

Adding accessory data to the RF model caused the scores to drop (macro F1 had the biggest drop from 0.76 to 0.71). The accessory data (wires and sidewalk conditions) were binary features, and there was a lot of information collapse in their construction. It is possible that these features no longer carried much meaningful information.

Adding the PLUTO features increased the macro F1 score of the RF Classifier from 0.76 to 0.92.[4] This jump in score cannot be attributed to the dramatically fewer data points in the dataset with PLUTO features: using those same data points without including PLUTO features resulted in a macro F1 of 0.29, much lower than the original score of 0.76 on the larger dataset. Figure 5 shows the final validation results of the best two classifiers.

Feature Importance

The most salient features for the RF Classification are shown in Figure 6. Though the accessory features did boost the scores slightly with the inclusion of PLUTO data, they are still the two least important features. The "wire presence" feature, which had the most information loss of all the features in the matching process, is last in importance. The "species" feature may

[4]Logistic Regression failed to converge when run on the dataset with PLUTO features, which could be due to the smaller size of the dataset.

	Model	*Accuracy*	*Macro F1*	*Weighted F1*
1,008,859 Samples	Linear Regressor	0.01 (R2 score)	-	-
	RF Regressor	0.42 (R2 score)	-	-
	Logistic Regressor	0.19	0.12	0.12
	RF Classifier	0.81	0.76	0.81
	RF Classifier + Accessories	0.79	0.71	0.79
3,091 Samples	RF Classifier + Accessories	0.57	0.29	0.56
	RF Classifier + PLUTO	0.92	0.89	0.92
	RF Classifier + Accessories + PLUTO	**0.94**	**0.92**	**0.94**
	GB Classifier + Accessories + PLUTO	**0.94**	**0.92**	**0.94**
	SVM + Accessories + PLUTO	0.84	0.64	0.88

Figure 4: Results for models trained on upsampled data (one run).

have faced a reduction in helpfulness since it required a numerical encoding strategy that also resulted in information loss.[3]

Latitude bin is the most important feature, followed by the average number of floors of the buildings in the bin, and longitude bin. Removing the latitude and longitude bin features yielded almost the same scores as with them included, however removing zip code began to show a degradation in scores, signaling that at least some location data is necessary to achieve high quality predictions. The fact that specific location is significant in inferring information about city greenery is a well-established concept [5], and it is likely that relevant neighborhood-level socioeconomic or physical features are being represented by location in this dataset. Likewise, the number of floors in the buildings is probably encoding some of this information as well. The number of floors could also be indicative of the amount of sunlight or wind the trees in the vicinity receive. Future work could include some analysis on whether there is a positive or negative correlation between building size and tree health.

The "number of floors" feature probably does not suffer from as much information collapse as land use category or other potential features when averaged across buildings in a bin, since building sizes tend to be relatively similar in adjacent buildings. The land use average may suffer from more information collapse, which could indicate its lower predictive power. Land use is a mixture between a categorical and continuous feature, with discrete land use codes residing on a scale that is somewhat continuous in meaning (one and two family buildings coded as 1, multi-family walk ups coded as 2, multi-family elevator buildings coded as 3, etc),

Model	*Accuracy*	*Macro F1*	*Weighted F1*
RF Classifier	**0.917 (0.010)**	**0.851 (0.100)**	**0.917 (0.010)**
GB Classifier	0.911 (0.008)	0.843 (0.101)	0.911 (0.008)

Figure 5: Results for best models averaged over five different train/test splits with all features included.

although this pattern is not robust throughout the all the categories (see Figure 7 for all land use categories). There were not many instances where these features needed to be averaged, so the effects of averaging may be slight, however future work could include exploring different ways of aggregating "number of floors" and "land use," as well as tree health, including testing a majority vote strategy rather than averaging.

Feature	*Importance*
Latitude Bin	0.2093
Average number of floors in buildings	0.1913
Longitude Bin	0.1516
Zip code	0.1499
Species	0.1293
Average land use code	0.0810
Tree diameter	0.0736
Sidewalk condition	0.0106
Wire presence	0.0034

Figure 6: Gini importance of all features in the Random Forest Classification model.

Conclusion

With this research I have shown that a simple model trained on publicly available data can predict the average health of NYC street trees with 90% accuracy. This research can be used to give insight into the street trees that might need more attention in the future. Theoretically the models could run on the feature data from the 2015 census (or the upcoming 2025 census) to predict where street trees will be in poor health in the future, and proactive measures could be taken to improve the health of those trees.

Perhaps the biggest limitation of this study, and a good area for future exploration, is the binning and averaging over features and tree health status that was necessary to join past

VALUE	DESCRIPTION
01	One & Two Family Buildings
02	Multi-Family Walk-Up Buildings
03	Multi-Family Elevator Buildings
04	Mixed Residential & Commercial Buildings
05	Commercial & Office Buildings
06	Industrial & Manufacturing
07	Transportation & Utility
08	Public Facilities & Institutions
09	Open Space & Outdoor Recreation
10	Parking Facilities
11	Vacant Land

Figure 7: Land use categories from the PLUTO Data Dictionary [6].

data with future data. Because of the way the information is collected by the census takers, it is currently impossible to do this work on the basis of individual trees, which makes the tree health predictions less meaningful. However, in future tree censuses, perhaps it will be possible to mark more accurate latitude and longitude coordinates so that trees can be identified on an individual basis, allowing predictions of their individual health rather than aggregate health. Future work in this space could also include using more publicly available NYC datasets as extra features (ex: air quality, temperature) to make more informed tree health predictions.

References

[1] *1995 Street Tree Census | NYC Open Data*. URL: https://data.cityofnewyork.us/Environment/1995-Street-Tree-Census/kyad-zm4j (visited on 05/01/2022).

[2] *2005 Street Tree Census | NYC Open Data*. URL: https://data.cityofnewyork.us/Environment/2005-Street-Tree-Census/29bw-z7pj (visited on 05/01/2022).

[3] *2005 Street Tree Census Data Guide and Field Definitions*. URL: http://www.geography.hunter.cuny.edu/~wenge/EDS_2021F/data/NYCtree/NewYorkCity_StreetTreeCensus2005_DataDescription.pdf.

[4] *2015 Street Tree Census - Tree Data | NYC Open Data*. URL: https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh (visited on 05/01/2022).

[5] Lorien Nesbitt et al. "Who has access to urban vegetation? A spatial analysis of distributional green equity in 10 US cities". en. In: *Landscape and Urban Planning* 181 (Jan. 2019), pp. 51–79. ISSN: 01692046. DOI: 10.1016/j.landurbplan.2018.08.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169204618307710 (visited on 05/01/2022).

[6] *PLUTO and MapPLUTO*. URL: https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page (visited on 05/01/2022).

[7] *Primary Land Use Tax Lot Output (PLUTO) | NYC Open Data*. URL: https://data.cityofnewyork.us/City-Government/Primary-Land-Use-Tax-Lot-Output-PLUTO-/64uk-42ks (visited on 05/01/2022).

[8] B. Skiera and G. Moll. "The sad state of city trees." en. In: *undefined* (1992). URL: https://www.semanticscholar.org/paper/The-sad-state-of-city-trees.-Skiera-Moll/7677e25c83b865c1335f3541eed7247a3f05225f (visited on 05/01/2022).

[9] *Street Tree Planting : NYC Parks*. URL: https://www.nycgovparks.org/trees/street-tree-planting (visited on 05/01/2022).

[10] *Young Street Tree Mortality : NYC Parks*. URL: https://www.nycgovparks.org/trees/ystm (visited on 05/01/2022).