DS-GA 1001 Intro to Data Science
December 19, 2022

An Analysis of NYC Street Trees

Team: Lorax (we speak for the trees)

Authors & Contributions

Yichen Shi (ys5538) - Inference
Fiona Chow (fc1132) - Prediction
Christine Gao (cg3278) - Prediction
Sicun Chen (sc9904) - Classification
Xinyue Ma (xm618) - Classification

1. Introduction

Urban trees provide a multitude of benefits to city residents. In addition to the environmental benefits trees offer such as lowering the temperature in urban areas and reducing flooding, urban trees also contribute to physical and economic health by creating walk appeal and increasing property value. Street trees are unique in that they are overseen by public municipal authorities, but they are difficult to manage because they are dispersed within the city in close proximity to private properties. Additionally, recent studies have shown significant associations between socioeconomic factors and street trees in urban neighborhoods. This report will attempt to reveal the associations between tree characteristics and geospatial properties, socioeconomic characteristics of the regions the trees are located in, and also motivate efforts to maintain healthy trees within NYC.

We aim to answer the following questions using the NYC street tree census dataset:

1. What tree features are related to tree health? Do different demographic groups show different tree statistics?
2. Can we use different demographic characteristics to predict tree density in the respective Neighborhood Tabulation Areas (NTAs)?
3. For the 10 species that have the most data points, can we predict the health status of a tree based on the features included in this dataset? Is one feature more important/more useful in predicting the health status for one species than the other?

Data

Our dataset is the NYC street tree census from 2015. The data contains 683,788 records of unique trees documented by volunteers and staff organized by NYC Parks and Recreation and partner organizations. The tree data has 41 columns that include data regarding **tree species, diameter, and health.** Additional variables include geocoded information regarding location of each tree. The parameters selected within this report include:

- curb_loc - location of tree in relationship to the curb (*OnCurb, OffsetFromCurb*)
- guards - indicate whether a guard is present and if it was helpful/harmful (*Harmful, Helpful, None, Unsure*)
- root_stone/root_grate/root_other - presence of a root problem caused by paving stones, metal grates in the tree bed, or other root problems (*Yes/No*)
- trunk_wire/trnk_light/trnk_other - presence of a trunk problem caused by wires wrapped around the trunk, or other trunk problems (*Yes/No*)
- brcnh_ligh/brnch_othe - presence of a branch problem caused by lights or wires wrapped in the branch, lighting installed on the tree, or other trunk problems (*Yes/No*)

- borocode/boroname - Code/name of borough in which the tree is located (*ex: 1 = Manhattan, 2 = Bronx)*
- nta/nta_name - NTA code corresponding to the neighborhood tabulation area from the 2010 US Census
- latitude/longitude - latitude and longitude in decimal degrees
- health - tree health status (*Good, Fair, Poor*)
- tree_dbh - diameter of the tree in inches, measured at approximately 54 inches above the ground (*Integer*).
- spc_common - common name for species (*ex. cherry*)
- user_type - describes the type of user who collected the data (*Volunteer, TreesCount Staff, NYC Parks Staff*)

For our analysis, we utilized additional demographic data joined on the street tree census data set by Neighborhood Tabulation Area (NTA) codes, provided by the City of New York Urban Planning census geography. We used multi-year estimates to increase statistical reliability for small population subgroups.

Missing Data/Extreme Values Handling & Data Preprocessing

- In our inference analysis, we cleaned the tree census data by NAs removal and manual outlier rejection. For the additional NYC demographic data, we convert its census measurement into the measurement at the borough tract level (boro_ct) as the tree data. Then we created a combined data by aggregating the tree data on the same boro_ct and obtained the new aggregated tree features together with the demographic factors. The combined data has around 1.8k rows.
- For our prediction question, we selected trees from the census data by accounting only for the live trees, then normalizing by removing any extreme outliers outside the 98th percentile. Variables were selected based on existing literature (demographic characteristics that most impact tree density in urban areas).
- For the classification question, we implemented row-wise removal for data points that contained missing values. We filtered for only alive trees that belong to the top 10 species, removed categorical variables with more than 10 levels, binary encoded the health variable into 1*(Good)* and 0 *(Not good*), and grouped the remaining data by species. The next step was to implement Factor Analysis of Mixed Data (FAMD) for dimensionality reduction and Lasso regression for feature selection, both detailed in later sections.

2. Inference Question

We combine our census data with additional NYC demographic data and consider two parts. First, which tree features are related to tree health? Second, as the demographic dataset shares the same variable borough census tract (boro_ct) as the tree data, both datasets are measured at the borough census tract level. For the aggregated data, we are interested in whether different demographic groups show different tree statistics (i.e. tree counts).

Question 1: Is there evidence of a relationship between the curb location (on curb/offset from curb) and sidewalk damage (damage/no damage)?

Since the two variables are categorical, we use the chi-square test. The null hypothesis is that there is no relationship between each variable and they are independent, and the alternative implies there is a relationship. We set the significance level to be 0.05, and check the assumptions that category levels are mutually exclusive and all expected values are greater than 5. We achieved a chi-square score of 2,558 with a p-value of 0.00. Based on the test statistics, we conclude that **curb location and sidewalk damage are dependent**. To account for large data size and high probability of falsely rejecting the null hypothesis, we calculate relative risk. The ratio of proportion of on curb trees showing sidewalk damage versus off curb trees showing sidewalk damage. The ratio is 2, so we conclude there is a practical effect and on curb trees are twice as likely to show sidewalk damage**.**

Question 2: Is there a relationship between the tree health and curb locations, conditioned on sidewalk damage?

Again, we used the chi-square test for categorical variables. For testing the relationship of the tree health and the curb locations, the chi-square score is 14.95 and the p-value is 0. However, the relative risk is almost 1, which indicates that these two variables might not have a practical relationship. When conditioned on no sidewalk damage, the p-value is 0.6, which fails to reject the null. When conditioned on the existence of sidewalk damage, the chi-square score is 80 and relative risk is 1.3. Thus, **we can draw no conclusion about the relationship between tree health and curb location**. More data or features need to be considered.

Question 3: Does there exist a relationship between tree statistics and demographic factors?

The tree features include total tree counts, number of distinct tree species, and percentage of healthy trees. The demographic features include the percentage under poverty, unemployment rate, and income per capita. All variables are represented by 1684 unique rows as a group statistic under each *boro_ct*. To investigate how the tree statistics vary (i.e. between low and high percentage of below poverty*)*, we conduct a median split for each demographic feature and derive tree stats for each group to test for differences.

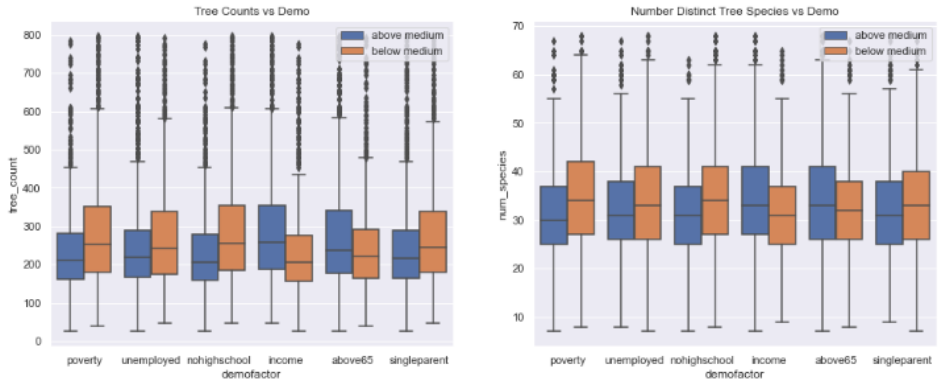

Figure 1: Boxplots of demographic factors corresponding to tree counts and number of distinct species

Fig. 1 uses boxplots to show the quantile distribution of the tree numbers for above and below median demographic groups. We observed that groups that tend to have more well-off statistics (i.e. lower unemployment, higher per capita income) have higher median tree counts and species diversity than their counterparts. Next, to verify our intuition with the hypothesis test on each of the six demographic variables, we use the Mann Whitney U test to compare medians. The null hypothesis is that the two groups come from the same population, the alternative being they are from different populations. The result of our test shows the U-scores for all demographic factors are very large with p-values of 0.00. Thus, we reject our null hypotheses and are confident the groups are from different populations and have different medians. Thus, we conclude that **some groups have higher tree counts and greater species variety than others.**

In addition, We analyze how the demographic variables relate to the percentage of good, fair and poor tree health and also how they are related to the percentage of sidewalk damage. Since the values in each group are all percentages, they are continuous and we use KS test to compare their distribution. As a result, the p-values are all not significant and we fail to reject the null hypothesis that the below median demographic group and the above median group have the same distribution. Therefore, **regarding tree health and sidewalk damage, we don't find significant relationships between the demographic factors.**

Now that we have shown that there is a significant difference in medians between tree counts and demographic groups, we will attempt to predict tree density with demographic characteristics in the next section.

3. Prediction Question

Using demographic characteristics in New York City neighborhoods, can we predict the street tree density in the area adjusting for confounders such as tree diameter and population density?

We are predicting street tree density as it gives insight into quality of life in an area. Street trees provide shade from the urban heat which is increasingly important with climate change. They also provide numerous ecological benefits to the community; an estimated 130 million dollars in ecological financial benefits each year through stormwater intercepted, energy conserved and air pollutants removed (City of New York, 2022).

Tree diameter is a confound because larger tree trunks are grown in the outer boroughs where there is land area, thus increasing tree density. Similarly, population density is a confound as more densely populated areas have more trees to promote ecological and economic benefits of street trees.

Approach

The demographic metrics span across race, education, income and wealth. We selected five socioeconomic variables: percent African American, percent of people with a bachelor's degree or higher, median household income, median estimated house value, and percent of owner occupied housing units. We chose the percent of variables where relevant to account for different sized neighborhoods. Similarly, we chose the median over the mean to account for any extreme outliers.

We selected trees from the census data by accounting only for the live trees (removing stumps and dead trees), then normalizing by removing any extreme outliers outside the 98th percentile. We computed tree density by counting the number of live trees per square feet of each NTA.

Methods

We started with a multiple linear regression model. We partitioned the data using an 80/20 split and trained our model to make predictions using the test set. To evaluate the accuracy of the model, we computed the coefficient of determination $R^2 = 0.620$. We chose R^2 to interpret our results over the mean squared error as the actual tree density proportion is very small and a very small mean squared error does not provide much insight.

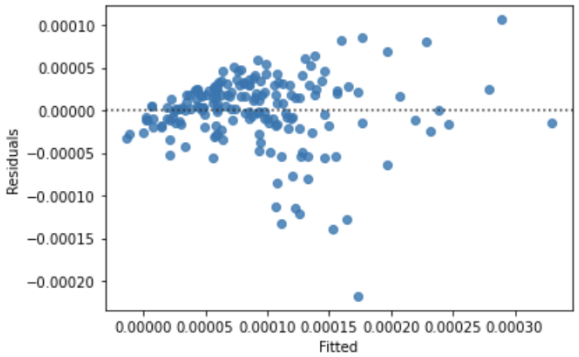

Figure 2: Residual Plot

The residuals in **Fig. 2** show that the points are randomly distributed around 0, with the exception of a few outliers. However, since the residual plot shows no curvature, we can assume linearity and will stick to our multiple linear regression model. We further evaluate our model by conducting a k-fold cross validation with a parameter of k=10. We computed the mean test score on the k-folds which yielded a value of 0.454. Since the accuracy of our cross-validated model is roughly the same as our original R^2, we conclude that the model is not overfitting and is the best fit for predicting tree density.

Now that we have shown the associations between socioeconomic characteristics of the regions the trees are located in, next we will seek to find the relationship between tree species and health of trees.

4. Classification Question

For the 10 species that have the most data, can we predict the health status of a tree based on the features included in this dataset? Is one feature more important/more useful in predicting the health status for one species than the other?

Approach

Our approach is to first find a way to reduce the number of features in our data and then find a random forest model to predict the health status of each of the top 10 species. From there, we evaluate our model performance and conclude whether specific features are more important than others. During exploratory data analysis, we noticed that the target class was imbalanced (~80% of the data were labeled "1" and the rest were labeled "0"), so we selected to implement stratified train-test split and F1 score as our scoring metric to avoid misleading results.

Clustering (Dimensionality Reduction)

Due to a large number of data points and features as well as the dominance of categorical features in our data, we need to find a way to significantly reduce our features to be able to build classification models more efficiently. We first conducted a FAMD to explore the presence of collinearity and if a reduced dataset to a small dimension would be ideal. By leveraging the 'prince' package in python, we computed the explained variance for the first 20 components as shown in **Fig. 3** below. The top 2 and 6 components explain about 30% and 60% of the variance in the data respectively. For visualization, we randomly sampled 100 data points from the data and plotted their first 2 combined components in **Fig. 3**. A large area of overlap explains the relatively low explained variance of the first 2 components from the small sample. Due to the low combined variance explained by the first few components and difficulty of interpreting each combined component, we decided to perform feature selection instead.

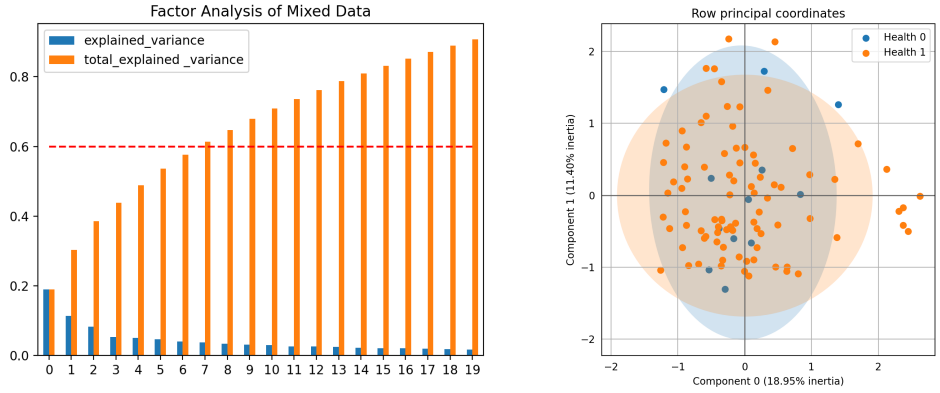

Figure 3: Individual/cumulative sum of explained variance (left) and dimensionally reduced sample data (right)

Feature Selection (Lasso Logistic Regression)

When performing feature selection with Lasso Regression, we strive to find a balance between the level of regularization ('C') and model performance (F1 score). After splitting the data in a stratified way, we used the training set to tune the model while monitoring its performance on the test set. By setting the penalty argument to l1, we found that the lasso logistic regression with C=10 degree of regularization produces an optimal score of 0.51. From this starting point, we gradually increased the level of regularization and monitored the decrease in F-1 score. When C is reduced to 0.01, we still have a 0.505 F1 score on the test set. We stopped the process as the test set F1 score reached 0.5. At this level of regularization, the output of coefficients is sparse and suitable for feature selection. We selected 12 features from the original (after preprocessing) 30 features whose coefficient is non-zero.

Classification Models

With a subset of features selected, the next step was to train models for each of the top 10 species. We chose to use random forest and included the following hyperparameters: criterion (ways to measure the quality of a split), n_estimators (number of trees in the forest), max_depth (maximum depth of each tree), min_samples_leaf (minimum fraction of samples required at a leaf node), and max_features (number of features to consider at each split). For each of the 10 species, we subsetted the trees in that species and partitioned the subset dataset using a stratified 80/20 train/test split, performed hyperparameter tuning, 3-fold cross validation and fit the best model on the training set. Using the best refitted model, we made predictions using the test set, extracted the model's feature importance and recorded its performance.

Model Performances

After making predictions using the testing set, we incorporated the confusion matrix, ROC curve (AUC score) and F1 score to assess the quality of the models. Based on **Table 1** and **Fig. 4**, the AUC scores of all models are between 0.6 and 0.7 which means the quality of the models is moderately good, while the F1 scores are mostly below 0.5 indicating the model is not performing very well. Overall, there is still room for improvement. Finally, we also noticed that models optimized for F1 scores tend to favor simpler models, and only the model for 'Norway maple' used different hyperparameters compared to others. The model for 'Norway maple' is also the only model which scored above 0.5 in F1 score.

	criterion	max_depth	max_features	min_samples_leaf	n_estimators	F1	AUC
London planetree	entropy	3	2	0.02	100	0.457259	0.680035
honeylocust	entropy	3	2	0.02	100	0.458909	0.643825
Callery pear	entropy	3	2	0.02	100	0.449248	0.628835
pin oak	entropy	3	2	0.02	100	0.461387	0.659842
Norway maple	gini	6	4	0.02	200	0.509068	0.638939
littleleaf linden	entropy	3	2	0.02	100	0.442246	0.638452
cherry	entropy	3	2	0.02	100	0.455804	0.598361
Japanese zelkova	entropy	3	2	0.02	100	0.463562	0.683370
ginkgo	entropy	3	2	0.02	100	0.448873	0.652575
Sophora	entropy	3	2	0.02	100	0.450178	0.636771

Table 1: Hyperparameters, F1 scores, AUC score of random forest models for the 10 species

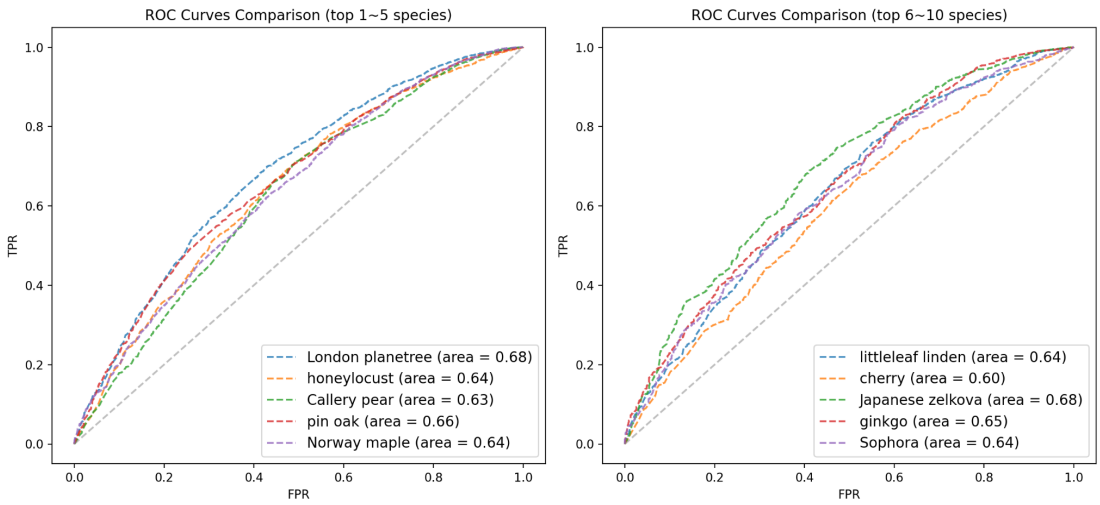

Figure 4: ROC curves of random forest models for the 10 species

Feature Importance

Regarding feature importance, we combined dummy variables that belong to the same variable and averaged the importance score since they have a higher chance of being selected. We can see that some features are definitely more important/more useful in predicting the health status for one species than the other based on **Fig. 5**. For example, the presence of other trunk problems (trnk_other) is important in predicting health status for Norway Maple, but not so important for Cherry.

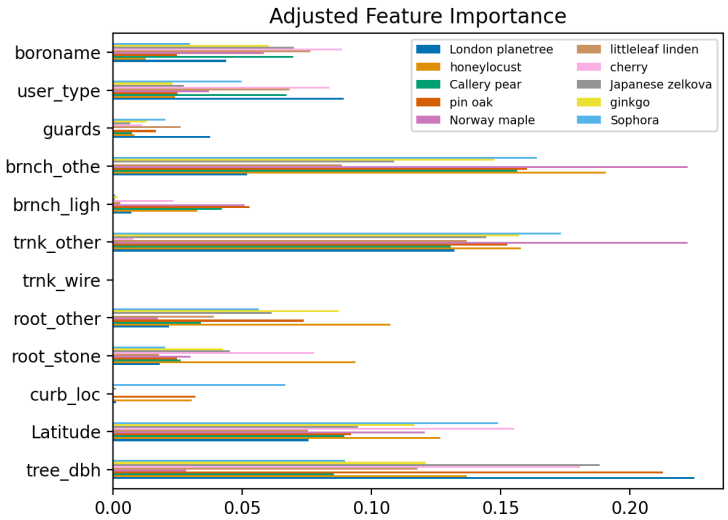

Figure 5: feature importance of random forest models for the 10 species

5. Conclusion & Discussion

From our initial inference, we found that on-curb trees are twice as likely to show sidewalk damage, while there is no relationship between tree health and curb location. There are inequalities in the tree counts and number of species for the demographic groups . Also, under the boro_ct level, the groups with higher percentages of under poverty, unemployment, low education levels and lower income are more likely to have less tree abundance and species variety. However, there's no clear relationship between the groups regarding the tree health and sidewalk damages.

We determined we can predict relative tree density in a given neighborhood by taking into account its demographic characteristics. When predicting health status for specific tree species, we were able to build random forest models to make moderately good predictions based on the existing features, but there is still room for improvement. We also found out that different features have different importance levels in predicting the health status of trees that belong to different species.

One assumption we made was that street tree data and demographic information of each neighborhood has not changed much since 2015 as the latest NYC street tree census data was published on NYC Open Data in 2015. Without any more recent data, we assumed the 2015 census data to reflect current statistics.

There are three major limitations of our analysis. The first one is regarding the availability of a more up-to-date dataset. Tree census data is updated once every ten years, making it difficult to include newly planted trees and demographic changes for a more recent analysis. The second limitation is Scikit Learn's inability to handle categorical features natively which forced us to use suboptimal strategies such as one-hot encoding. Some other implementations of random forest such as R's could natively handle categorical predictors without having to first transform them. The final limitation was target class imbalance, which we noticed during our classification analysis. Although we implemented stratified train-test split and F1 score as model measurement, a better approach such as over-sampling the minor class or under-sampling the major class could resolve this. We could also use other more computationally costly methods such as XGBoost.

With respect to the tree census data, we could not control the quality or assess reliability of the data sets. With higher quality data or a more recent inventory of the tree census and demographic information, we could provide a more accurate characterization of current conditions in NYC, thus improving any immediate decision making regarding resource allocation for improving equity and stewardship for street trees.

One thing we found interesting about the dataset was the distribution of tree species, notably that only 20 species make up 83% of NYC street trees. This could be an issue in the event a disease affects any of the species with the largest tree population. In addition to the severe ecological consequences, this poses a great economic challenge to regrow and replace the lost tree population.

Our analysis demonstrates that while street trees do correlate to higher quality of life, they reveal past inequalities that persist in the current distribution of street tree population. This has implications for urban planning when considering the species and location of future plantings, in addition to the aesthetics and considering equality of species distribution by region. We also highlight the inequitable distribution of NYC park resources, as neighborhoods with a higher level of inequity have lower tree density, making them more vulnerable to urban heat and with less access to the ecological benefits street trees provide.

With future census measurements and an improved understanding of the economic, ecological, and social impacts of street trees in the NYC metropolitan area, we hope that similar future analyses will inform local municipalities of the ever changing dynamics between street tree populations and human populations over time.

References

"2015 Street Tree Census" *NYC Open Data,* City of New York, 2022.
https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/pi5s-9p35
"NYC NTA Census Data" *NYC Open Data,* City of New York, 2022.
https://www.nyc.gov/site/planning/data-maps/open-data/census-download-metadata.page?tab=2
"CDC SVI 2016" *Centers for Disease Control and Prevention,* New York, 2022.
https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html
"Prince Library, Python" *Github*, Max Halford, 2022 https://github.com/MaxHalford/prince#readme

Code Appendices

2_Inference_P1_Processing.ipynb
2_Inference_P2_Hypothesis_Tests.ipynb
3_Prediction.ipynb
4_Classification_P1_Processing.ipynb
4_Classification_P2_Lasso_FAMD.ipynb
4_Classification_P3_Models.ipynb