NYC tree report

Roy Luo

Summary

The primary goal of this research is to find out how the geo-location in the form of latitude, longitude and other categorical variables interact/affect the status, health, and species classification of trees in NYC. The result shows the geo-location and numerous other categorical variables could have an effect in determining the likely status, health or species of a tree.

Data Description and Exploratory Data Analysis

The dataset is provided by NYC Open data and it collects information including tree species, diameter, perception of health and location information. 683,788 trees in NYC are recorded and each line of the dataset has 42 columns, including a unique tree ID number and other information regarding the tree such as tree species, diameter, and perception of health, etc. The data was conducted by volunteers and staff organized by NYC Parks & Recreation and partner organizations.

Since most of the columns are location-based, we will only select three variables that represent the location of a tree: longitude, latitude and borough name. Since borough gives a general block of the location while the other two variables are more specific. The other variables we will also be using are tree diameter(tree_dbh), unique signs a steward noticed for a tree(steward), sidewalk damage adjacent to a tree(sidewalk), location of tree bed in relation to the curb(curb_loc). And obviously, the aforementioned variables to classify: status, health, species.

Why experiment?

It is sometimes unclear for policy makers to determine where to plant trees to ensure the highest likelihood survival of the trees. To classify the likelihood of survival based on location could be rather useful in this case. Moreover, classification on the health of the tree could also help the local government determine the best location and how to plant the tree with relationship to sidewalk and curb. Finally, success in classifying between any two species could help researchers understand the clustering behavior and characteristics of these two species in comparison.

Exploratory Data Analysis

Data Preprocessing

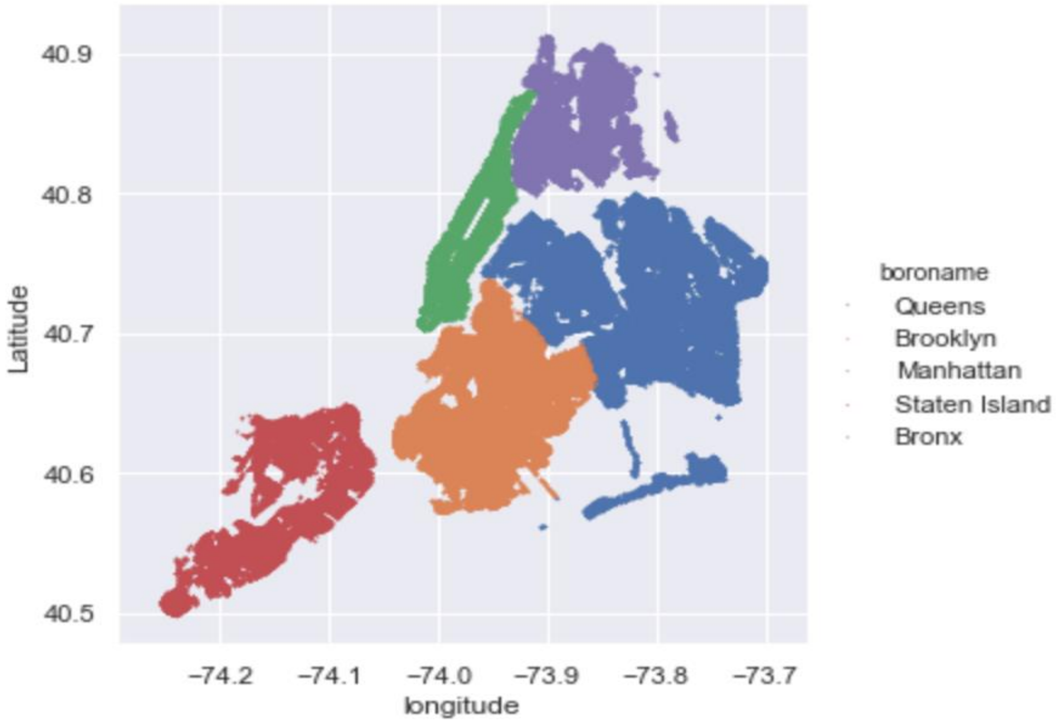
Categorical variables are hot encoded as indicator variables to make the model computations easier. All trees that are trunks are removed since they make up a small percentage of the dataset and it is more relevant to predict the probability of a tree is alive or dead given information of the tree. Moreover, since the health of a tree is only available when the tree is alive, only alive trees will be selected to predict the likely health of a tree.

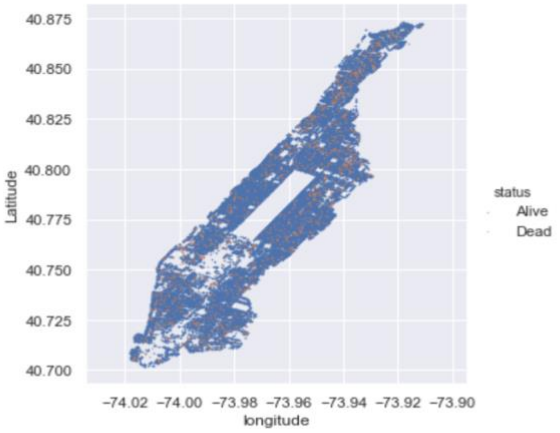
NYC tree report

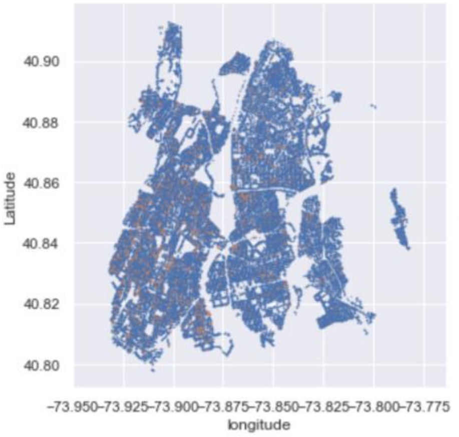
Roy Luo

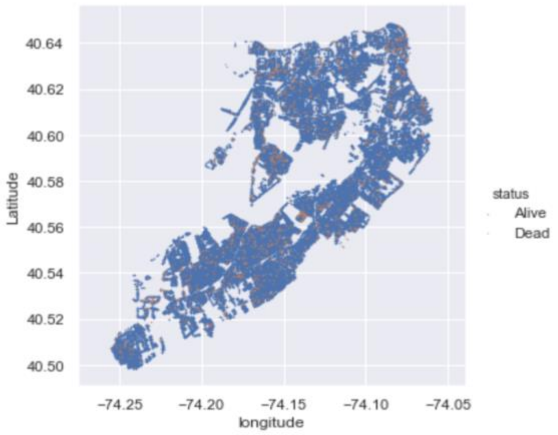
Scatterplots (Status)

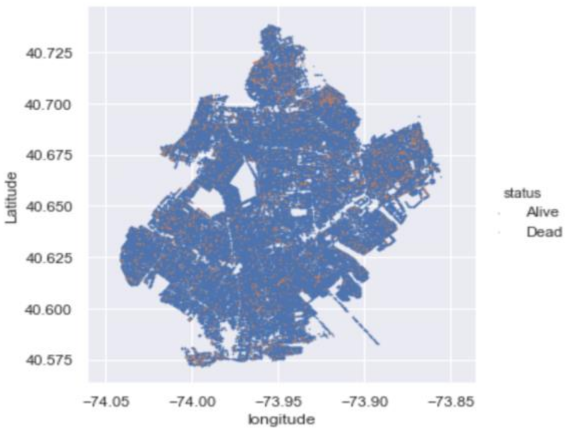
We first look at the overall geographical locations of the trees in the dataset which gives us a general idea of the locations of the trees. Notice that there are blocks missing in some major parks in NYC including the central park in Manhattan and prospect park in Brooklyn (Probably because there are too many and it is hard to map out those trees from major parks). Then a plot within each borough is plotted, with alive trees marked blue and dead trees marked red. It is noticeable that the alive and dead trees occur in clusters, it is likely that a KNN would create a nice decision boundary that effectively predicts the status of a tree. Additionally, each borough plot shows almost identical clustering behavior so it is likely unnecessary to train model on each borough separately.

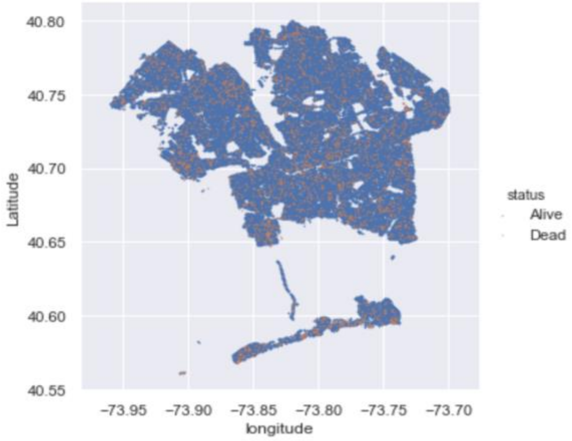






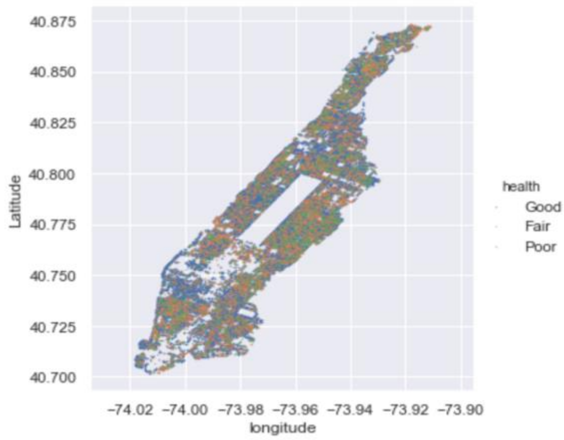


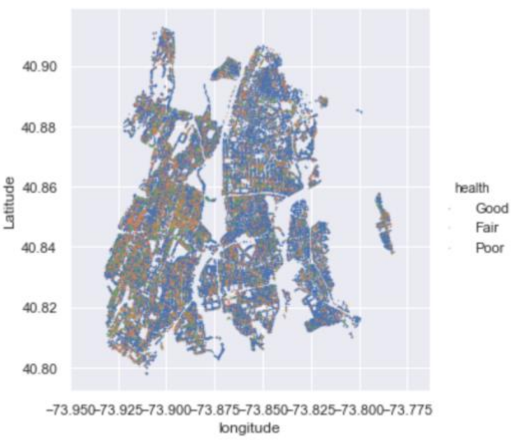


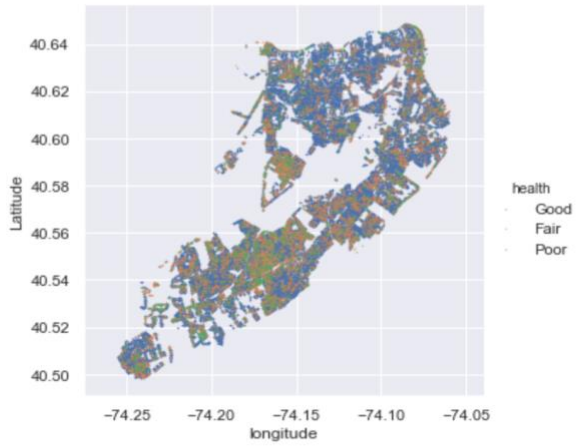


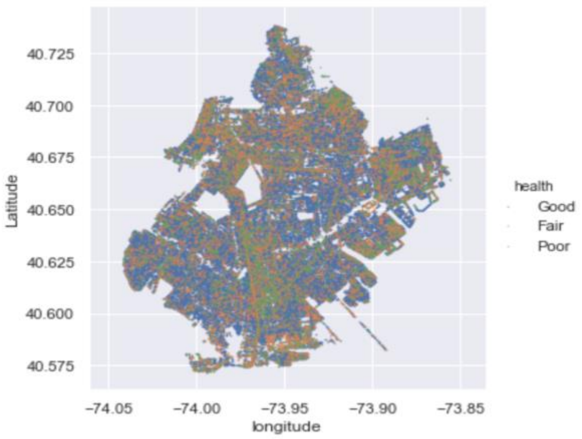
Scatterplots (Health)

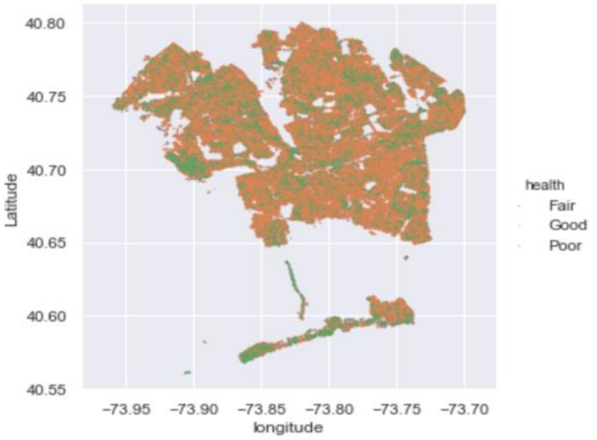
A parallel procedure is being done to check the geo-location effect on the perceived health of a tree. There is significant clustering behavior as well. Trees with similar conditions are clustered together in each borough. Unlike previous plots with the status of the trees, the behavior in health seems to be clustered differently in each borough, so it might be a good idea to separate each borough when training the model. This is likely due to a different local government policy and weather effect on tree health. It is interesting to find a higher likelihood of good health in Brooklyn.





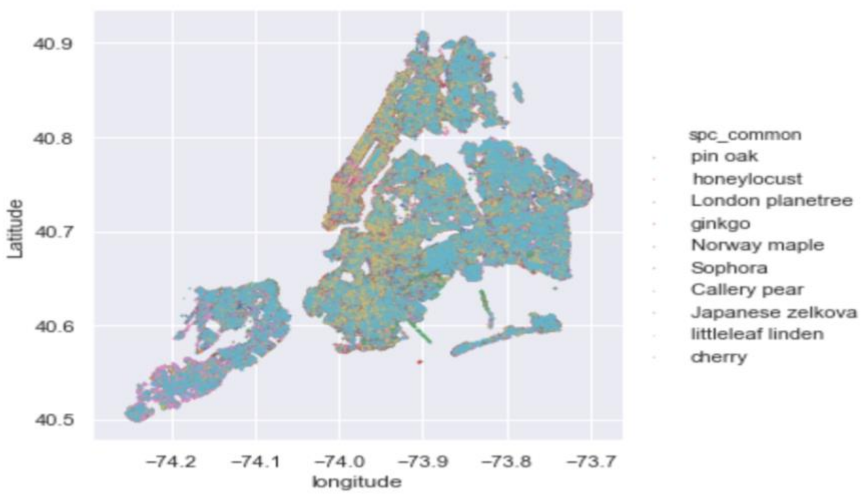






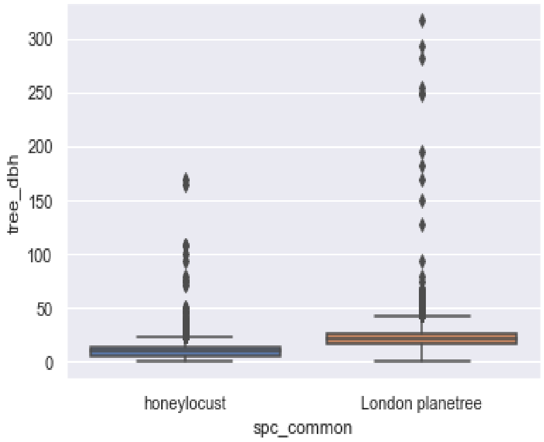
Scatterplot (Species)

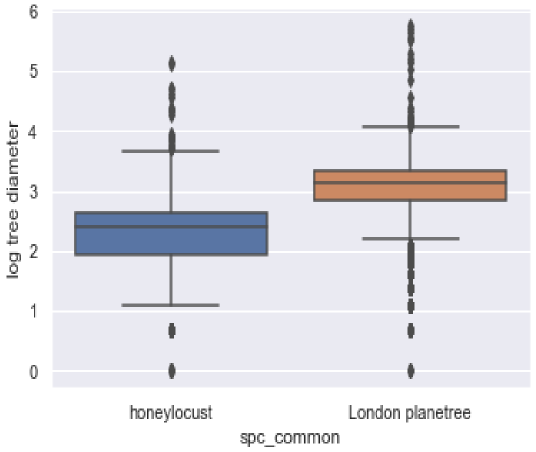
Finally, the top 10 most common species are plotted based on geo-location in the plot below. There is an even more clear clustering effect based on the locations. And we will likely use borough location as categorical variables in this classification instead of simply separating each model by borough since there will be numerous other variables.



Boxplots (Species)

In addition, the classification on the species will be done on the two most common species to test the quality of the classification models on the species. The two species that are selected are London Planetree & Honey locust. From the boxplots of tree diameter against the two species, there is a clear case of outliers and a logarithm transformation is conducted to make the numerical variable closer to normally distributed. And by looking at the distribution of tree diameter, it is easy to notice a significant difference between the two species. A simple two-sample t-test would also suggest a significant deviation between the two species with a p-value close to zero. It is safe to assume the two samples are different in tree diameter.





Model analysis

Status prediction (Alive/Dead)

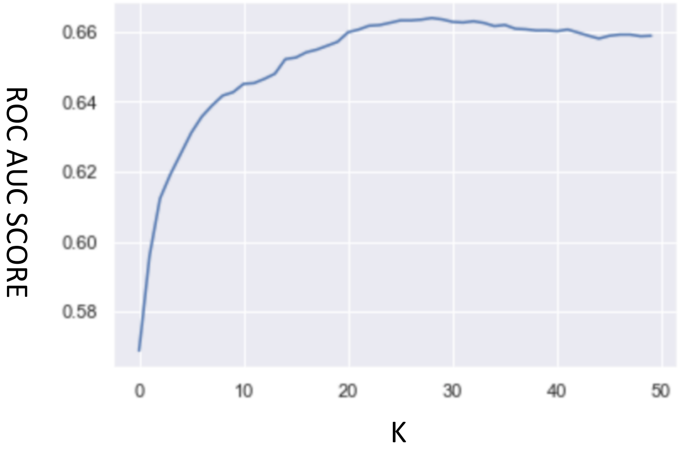
	FEATURE	ERROR	AUC
LOGISTIC REGRESSION	Geolocation	0.0207	0.5169
LOGISTIC REGRESSION	Geolocation, tree diameter	0.0207	0.7696
KNN (K=5)	Geolocation	0.0214	0.6252
KNN (K=5)	Geolocation, tree diameter	0.0217	0.6389
QDA	Geolocation, tree diameter	0.0207	0.7381
ADABOOST	Geolocation, tree diameter	0.0207	0.7730

Geolocation, in this case, represents the latitude and longitude. From the above outputs, it is interesting to note that the error rates are extremely similar. Since the base distributions between alive and dead trees are roughly 95%-5%, an error rate lower than 5% is considered performing better than random guessing. Another metric that is introduced is the ROC-AUC score. This metric shows the adaptability of a model in its response and sensitivity, so it has an important role in this classification case. Moreover, the logistic regression performs significantly better with the addition of tree diameter as a feature, while the KNN method improves very slightly with the addition of tree diameter. Overall, the best performing model is the Adaboost model with a ROC-AUC score of .773 which is significantly better than random guessing. The weak learner used in this ensemble method is a classification tree with a depth of 2 and max iteration of 200 with a learning rate of .5.

KNN performs well on geolocation alone which is expected due to the clustering behavior of trees that are close together. For instance, a location with bad soil or other location-based reasons could affect the survival rate of a tree significantly. A policymaker in charge of trees can access this model to predict the survival rate of a tree in a specific location to more effectively plan where to plant trees.

A search to find the best K-parameter was also conducted and the results from the plot below shows the best AUC score occurs roughly at k = 29 with a value of .663. This means that a tree's survival rate can be best predicted using the information on the surround 29 trees based on this specific model.

(Test of k in KNN)



Adaboost could be effective in predicting the status since all weak classifiers in this sample has a high chance of being independent of each other, which would make the Adaboost algorithm more powerful in predictions. And as expected, the Adaboost was the best performing model in this classification model.

Health prediction (only available when the tree is alive)

Geo-location and tree diameter features are used similar to when the status was predicted. Since health status includes three possible class: good, fair, poor; a weighted precision metric is being implemented to check the overall precision with a distribution weight given to each class.

	BOROUGH	WEIGHTED TEST AVERAGE PRECISION
KNN	Queens	0.74
	Brooklyn	0.75
	Manhattan	0.67
	Staten Island	0.76
	Bronx	0.75
ADABOOST	Queens	0.76
	Brooklyn	0.74
	Manhattan	0.67
	Staten Island	0.76
	Bronx	0.73
LOGISTIC	Queens	0.71
	Brooklyn	0.71
	Manhattan	0.66
	Staten Island	0.71
	Bronx	0.72

From the outputs above there is likely an effect in blocking specifically on Manhattan. Overall, Manhattan has the lowest prediction precision. Logistic regression performs slightly worse than the other two models, while KNN and Adaboost have almost identical prediction results. In this case, however, it is more elegant to apply KNN since it takes slightly less time to train while Adaboost could be computation heavy.

In theory, it is possible that more features would give a better result, but in reality, adding categorical variables 'steward', 'sidewalk', 'curb_loc' had almost no improvement on the model (actually there was a decrease in KNN weighted test precision). To explain this phenomenon, it is likely that the categorical variables add too much noise to the prediction results. Especially in the case of KNN where geo-location likely plays an important role in the prediction algorithm, adding too many features would only reduce the effectiveness of the model.

With KNN mode and combined with the model from status prediction earlier, a policymaker has the ability to determine the likelihood of survival and the likely health of a tree assuming its survival just based on the location and expected tree diameter. A tree diameter could be useful in these predictions due to the age representation in diameter and a tree of older age has a higher chance of survival.

Species Classification (London Planetree & Honey Locust)

The two most common species in NYC are the London Planetree and Honey Locust. From earlier EDA, we noticed a relationship between the tree diameter and the species type. Other categorical variables are also considered in this prediction to classify between the two species. Again, since this prediction is a binary classification, AUC will be closely examined. Moreover, the error rate will be an important metric since the species are balanced in total occurrences.

	ERROR	AUC
KNN	0.1657	0.9007
ADABOOST	0.1586	0.9177
LOGISTIC REGRESSION	0.2105	0.8709

From the above outputs, Adaboost clearly outperforms other models. KNN performs slightly worse than Adaboost but KNN significantly outperforms logistic regression which acts as a baseline in this case. The Adaboost shows a low prediction error rate and high flexibility in its probability predictions.

Furthermore, since Adaboost uses classification tree as its weak learner and borough is used as a categorical feature in this problem, Adaboost will automatically block each borough while KNN potentially has trouble minimizing other borough's effect when predicting near the edge of a borough. This could explain the slight discrepancy in KNN and Adaboost prediction results.

Conclusion:

Based on some model analysis and model comparison, we conclude that some variables specifically the geo-location and tree diameter are effective when predicting the status, perceived health, and any two species classification.

Since the KNN performs rather well in all three types of classification, it can be interpreted that future tree allocation should be around locations that have a higher number of alive and healthy trees. As for classification between species, it can be interpreted that same species tend to cluster in the same location.

There is an opportunity for future research on why some specific locations have a high survival rate and some locations have a low survival rate. For example, why is it so likely for a tree to survive in Brooklyn. And perhaps there are more reasons that are not presented in this dataset that could explain the survival rate of a tree. Some researchers could find other common characteristics in alive trees that helpful for future tree allocations.