Executive summary

This analysis was conducted on NYC's 2015-2016 tree census and available on NYC's open data website. In terms of model accuracy, Logistic regression was 76.8 % accurate and Tree classification was 97.9% accurate.

In terms of determining which boroughs were most likely to produce dead or stump trees, Logistic regression showed that Staten Island and Manhattan had the most effect while in tree classification none of the boroughs had any effect.

Background

This project looked at NYC open data for a tree census completed by the "NYC Parks & Recreation and partner organizations" in 2015 and was "conducted by volunteers and staff organized by NYC Parks & Recreation and partner organizations."[1]

The census brought 2,241 volunteers "together in the largest participatory municipal urban forestry project in United States history". In terms of technology the census used "innovative geospatial technology and a strong quality review process has yielded an exceptionally accurate inventory of street trees."[2]

The data contains 683,788 records where each record is an individual tree. Approximately 95% of the tree were designated "Alive" while the remaining 5% were "Dead" (2%) or "Stump" (3%) trees. Additionally, there were 45 variables including tree dimensions, geographic indicators, and health/problem indicators. The response variable used was called 'Status' and had three values: "Alive", "Stump" or "Dead".

[1] https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh

[2] https://www.nycgovparks.org/trees/treescount

Research purpose

The target audience for this report includes data analyst, environmentalists, tree advocates and those in the public sector.

The project attempts to answer two questions:

- 1st Question: Which model will best accurately predict the health of NYC trees and what variables will contribute significantly to that prediction?
- 2nd Question: Which NYC borough are mostly like to produce dead or stump trees?

Several statistical models were considered such as K – Nearest Neighbors, Linear Regression, Logistic Regression, Classification Tree, Bagging and Random Forest. Ultimately Logistic Regression and Classification Tree were used.

Data preparation

> *"A common rule of thumb is that 80% of the project is data preparation."*[3]

Perhaps not 80% in this case but a fair amount of time was devoted to preparing the data and/or determining which desired statistical models could work for the tree census data. KNN and linear regression were models both considered and rejected. The elimination of linear regression was easily determined since the response variable "Status" was categorical (with values 'Alive', 'Dead', 'Stump').

Eliminating KNN took more work especially since there was a desire to create KNN visualization maps that would have been helpful given the nature of the census data that included several geographical indicators. Several attempts were made such as selecting different subsets of the data and dealing with

[3] https://www.datascience-pm.com/crisp-dm-2/

"blank" data points. However, nearly all of 45 variables were categorical (with many different values) and made it difficult for the model apply Euclidean distance for the model to fit the data correctly.

To fit the classification tree model, the predictors had to be binary and two important variables had to be changed to accommodate this requirement. First the values for the response variable ("Status") were changed from the 'Alive', 'Dead', 'Stump', to 'Alive" and "DeadorStump". Additionally, to use the borough variable (with five values corresponding to Bronx, Brooklyn, Manhattan, Queens, Staten Island) it had to be changed to four separate binary dummy variables. For example, the newly made "Manhattan" variable had values 'Yes' or 'No'. "Yes" indicated that the tree was located in that Manhattan while "No" meant the tree was planted in one of the other four boroughs.

Note: in the oral presentation the logistic regression analysis used binary borough data. After reflected on the finding in the oral presentation, it was decided the logistic regression analysis should include only one variable ("borough") with five values – one for each borough.

In short, Logistic regression and tree classification were used for the data mining and 13 of the 45 variables were used. Variables were excluded either because of redundancy, relevance, or usability.

Model accuracy

Typically, the threshold value for logistic regression is 50%. When applied at this threshold the accuracy of the model was approximately 2%. This threshold had to account for whether a tree would be alive or dead/stumped and had to be the percentage of trees that were dead/stumped in the database or 4.62% (31,615 out of the 683,788 total). Based on that threshold the accuracy 76.8%. The model predicted better the number of trees dead/stumped than alive trees. For alive trees, the model predicted 76.3% correctly but for dead/stumped the prediction improved to 86.6%.

Using the 80-20 holdout resampling method - significant variables were: Tree Diameter (40.7%), borough-Staten Island (24%), borough-Manhattan (14.0%), borough-Brooklyn (13.7%), curb location – on curb (11.6%) and borough – Queens (11.3%). The percentages for each of these variables is the odds ratio. For example, holding all other variables constant a 1 unit increase in borough-Manhattan, the odds of the response variable equaling DeadorStump increases by 14.0%

As mentioned in the executive summary the overall accuracy was higher with the tree classification model (with optimal tree pruning set at 4 terminal nodes) was 97.89%. This time the model predicted better the number of alive than dead/stumped trees. For alive trees, the model predicted 99.97% correctly but for dead/stumped the prediction decreased to 55.85%. In this model the borough variables were not significant. Instead, tree diameter and root_stone ("root problems caused by metal grates"[4]) were found to be significant in the classification tree.

Conclusions

In terms of accuracy the tree classification is a clear winner. However, the information within it is less 'fruitful' than the logistic regression. Less fruitful in that it provided no information on possible borough influence. Also, it was striking that the tree classification's prediction changed significantly among alive or dead/stumped trees (99.9% vs 55.9%). Several unsuccessful attempts were made at running bagging and random forest models to gather more information on the tree results (perhaps due to structure of the classification data or computing power). A next step would be to revisit those models.

[4] Data dictionary found in the attachment section of: https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh
Taken from

The logistic regression model showed that Staten Island had the most influence on the response variable. Of the five boroughs, only the Bronx did not show as being significant. To answer this question would require additional research and may present interesting findings.

Figures

Of those designated “Alive” here is an approximate breakdown of those marked “Good”, “Fair”, “Poor”:

- Good trees: 528,850 (81%)
- Fair trees: 96,504 (15%)
- Poor trees: 26,818 (4%)

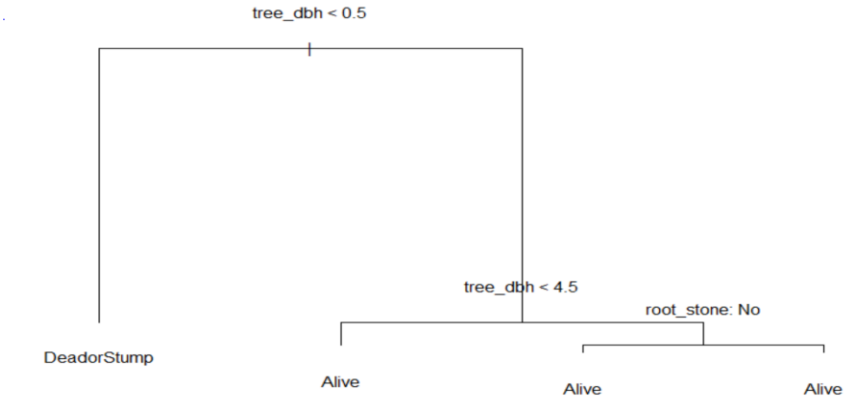
Logistic regression – confusion matrix

	true value				
prediction	alive		dead_stump		
alive	99,499	76.28%	847	13.41%	
dead_stump	30,945	23.72%	5,467	86.59%	
	130,444		6,314		136,758
0verall accurcy	76.75%				

Tree classification – confusion matrix

	true value				
prediction	alive		dead_stump		
alive	130,275	99.97%	2,844	44.15%	
dead_stump	42	0.03%	3,597	55.85%	
	130,317		6,441		136,758
0verall accurcy	97.89%				

Tree nodes



All 45 variables

status	spc_common	trunk_wire	community board	state
health	steward	trnk_light	borocode	latitude
tree_id	guards	trnk_other	borough	longitude
block_id	sidewalk	brch_light	cncldist	x_sp
created_at	user_type	brch_shoe	st_assem	y_sp
tree_dbh	problems	brch_other	st_senate	council district
stump_diam	root_stone	address	nta	census tract
curb_loc	root_grate	postcode	nta_name	bin
spc_latin	root_other	zip_city	boro_ct	bbl

Variables for logistic regression

```
[1] "status"     "tree_dbh"   "curb_loc"   "root_stone" "root_grate"
 [6] "root_other" "trunk_wire" "trnk_light" "trnk_other" "brch_light"
[11] "brch_shoe"  "brch_other" "borough"
```

Variables for tree classification

```
[1] "status"     "curb_loc"   "root_stone" "root_grate" "root_other"
 [6] "trunk_wire" "trnk_light" "trnk_other" "brch_light" "brch_shoe"
[11] "brch_other" "Manhattan"  "Bronx"      "Brooklyn"   "Queens"
[16] "StatIsland" "tree_dbh"
```