Verifying Volunteer Entries to the NYC Trees Census

Joshua Szymanowski

A look at the data

NYC OpenData

NYC Parks

TREES COUNT 2015

652,173 trees

444,390 professional (TreesCount and Parks staff) entries

207,776 volunteer entries

Health of trees (classification target):

Good *(82%)*, ***Fair*** *(14%)*, and ***Poor*** *(4%)*

Variables:

Tree diameter, species, number of stewards, quality of tree guards, root/trunk/branch problems

Location attributes:

Borough, neighborhood, community board, council district, state assembly, state senate

Latitude & longitude for **mapmaking**

Project Outline

1

Data Cleaning

Drop stumps and dead trees

Drop volunteer data

Turn data objects into numericals

2

EDA

Effects of location:
- Borough
- Neighborhood
- Political districts

Effect of species
Root, trunk, and branch problems
Tree guards
Tree stewards
Maps

3

Baseline Model

Fit a Random Forest Model with standard parameters

Interpret which features are most important and if any more EDA needs to be done

4

Feature Engineering

Distance of nearest tree using GeoPandas

Number of trees on same block

Create dummy variables

Use community board as neighborhood variable

5

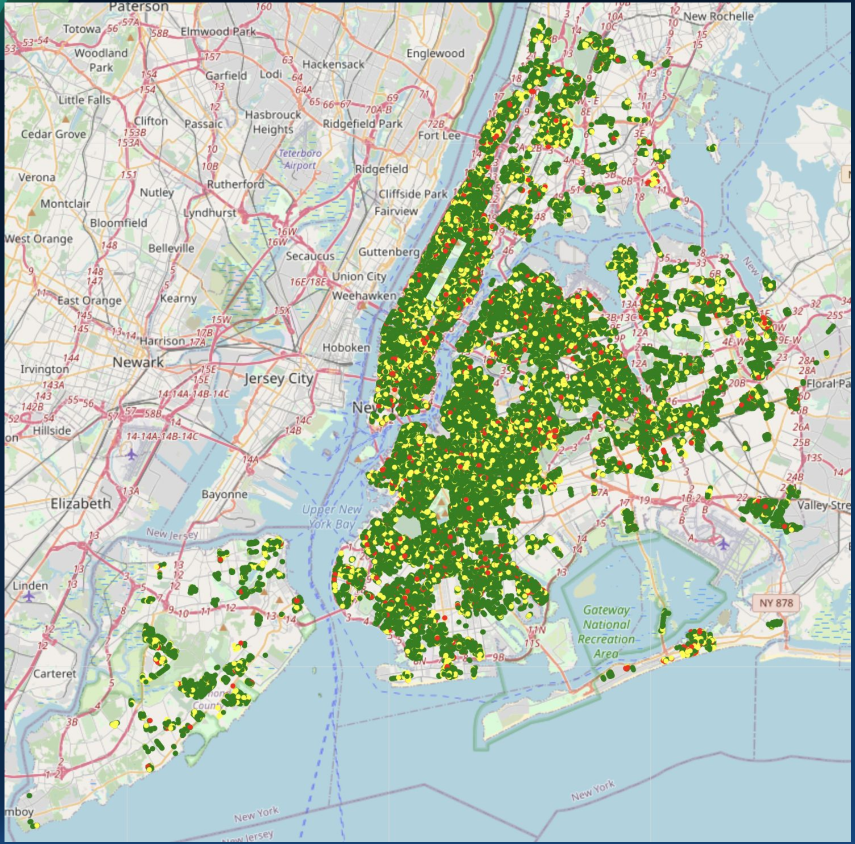
Final Model

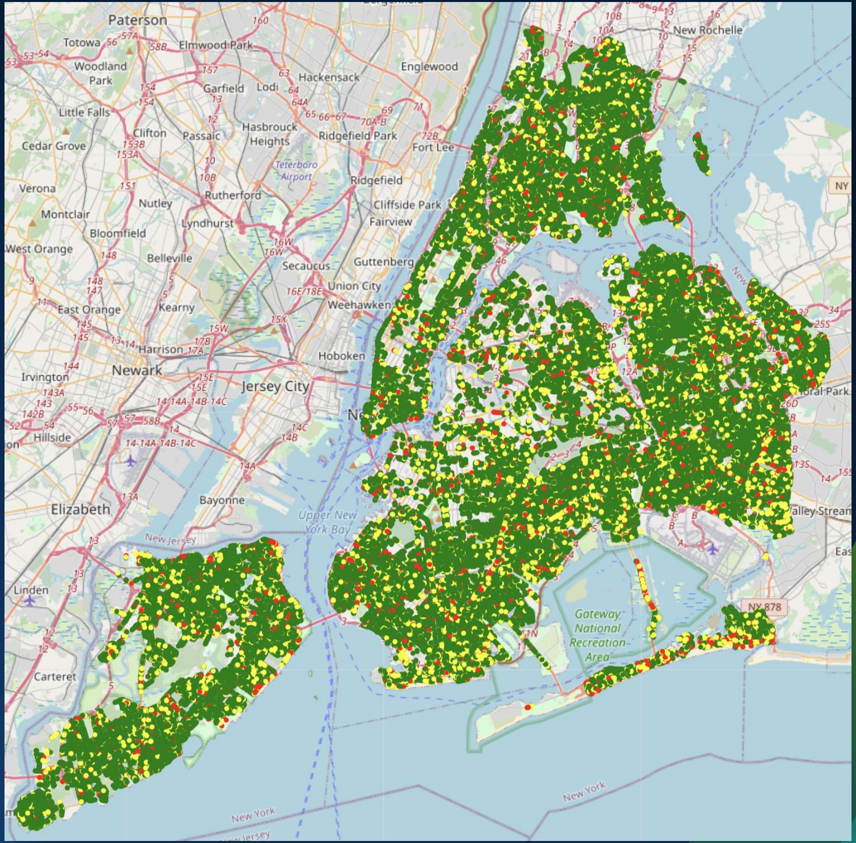
Random Forest with GridSearch

Naive Bayes with GridSearch

Create list of flagged trees (future goal)

Volunteer entries





Professional entries

🟩 Good
🟨 Fair
🟥 Poor

Goals

Recommend changes for next tree census

Recommend policies for trees planting and maintenance

Develop a model to verify the health status given by volunteers

Model preview

Vanilla random forest

Hyperparams: *class_weight=‘balanced’*

Accuracy: **54.0%**

Weighted F1: **61.1%**

Untuned random forest

Hyperparams: *class_weight=‘balanced’*

Accuracy: **84.6%**

Weighted F1: **81.7%**

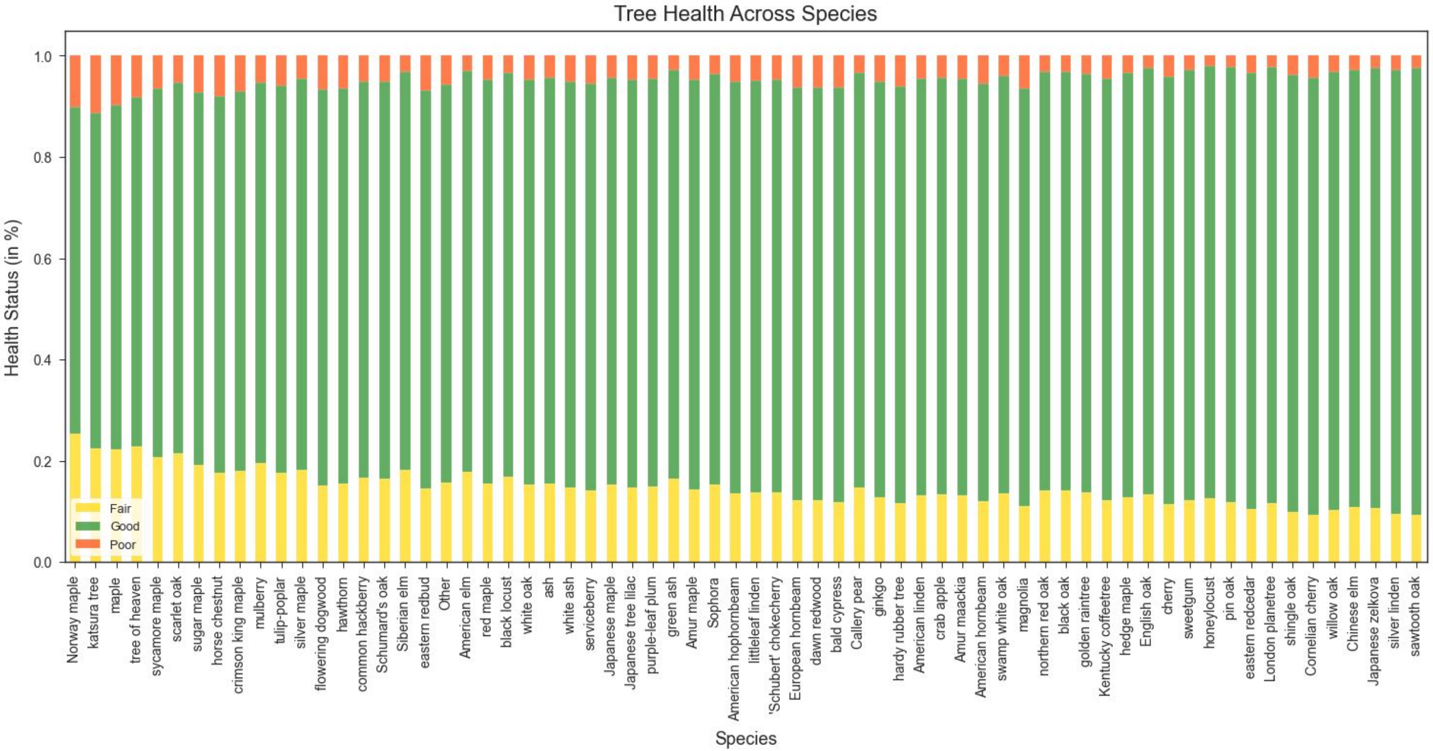
Tuned random forest

Accuracy: **77.1%**

Weighted F1: **78.2%**

Hyperparams:
max_features=11,
min_samples_split=11,
min_samples_leaf=2,
n_estimators=500,
class_weight=‘balanced’

Species matters



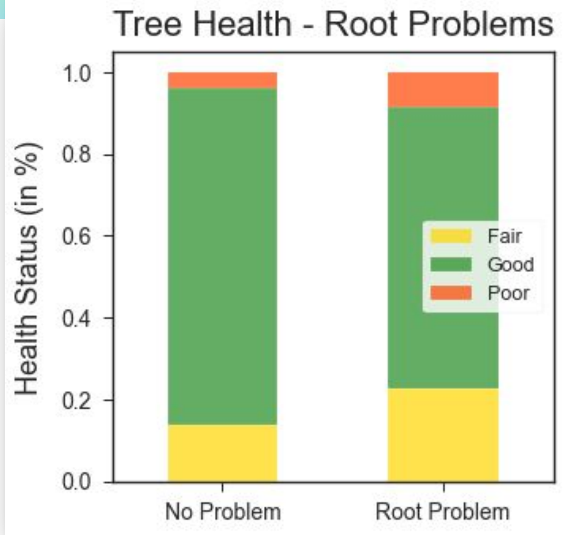
Top 3:

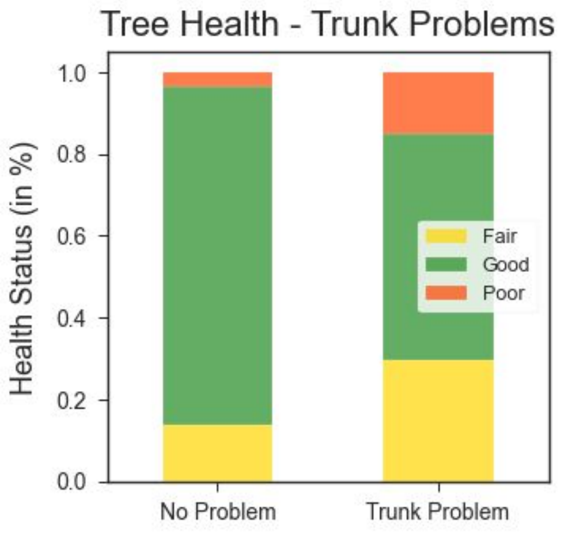
- **Norway maple**
- **Katsura tree**
- **Maple**

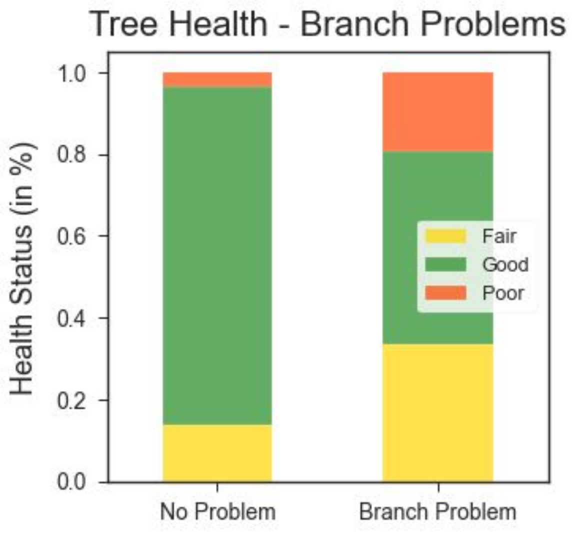
Bottom 3:

- **Sawtooth oak**
- **Silver linden**
- **Japanese zelkova**

Tree problems (and solutions)







The most important problems are listed as “Other”.

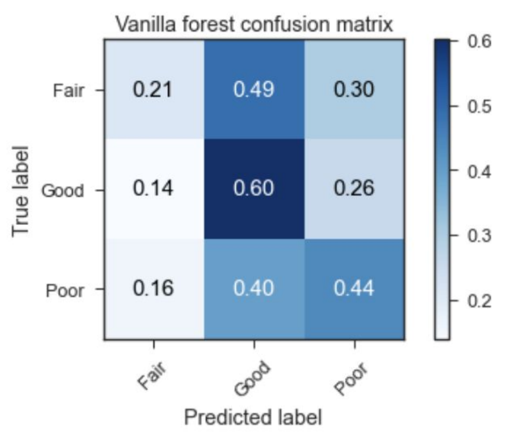
In future censuses, be more specific or have a notes column for each.

Solutions for healthier trees:

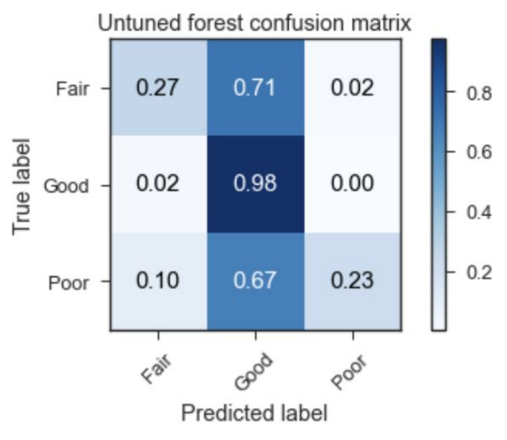
More regular maintenance of trees.

Farther reaching environmental protections, à la the plastic bag ban.

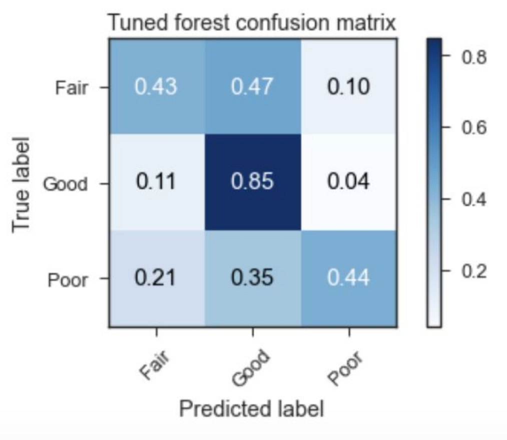
Confusion journey



Accuracy: **54.0%**
Weighted F1: **61.1%**



Accuracy: **84.6%**
Weighted F1: **81.7%**

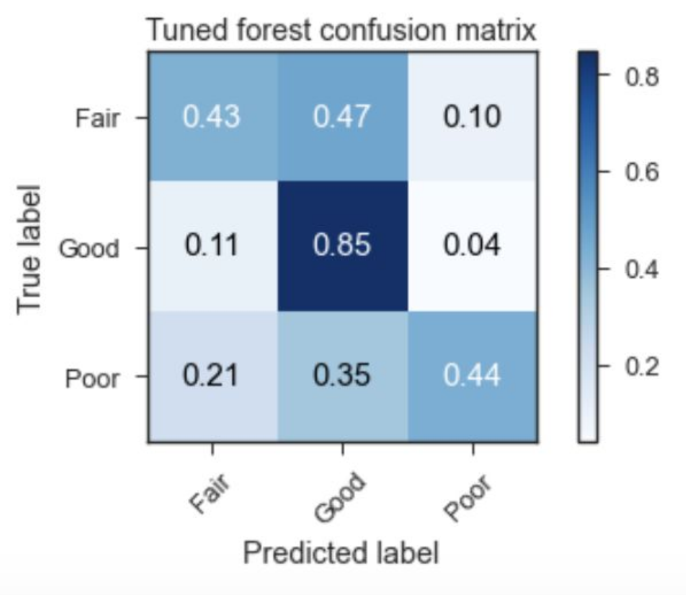


Accuracy: **76.8%**
Weighted F1: **78.0%**

Final model –
Random Forest

Top features (out of 147):

Distance to nearest tree	(15.8%)
Number of trees on block	(14.0%)
Tree diameter	(13.7%)
Species [Norway maple]	(2.1%)
Branch problems [other]	(2.1%)
Tree stewards [1-2]	(2.1%)
Trunk problems [other]	(2.1%)
Sidewalk damage	(1.9%)
Root problems [stone]	(1.3%)
Branch problems [light]	(1.2%)
On curb	(1.1%)
Species [London planetree]	(1.1%)
Species [Honeylocust]	(1.1%)
Tree guards [helpful]	(0.9%)
Root problems [other]	(0.9%)



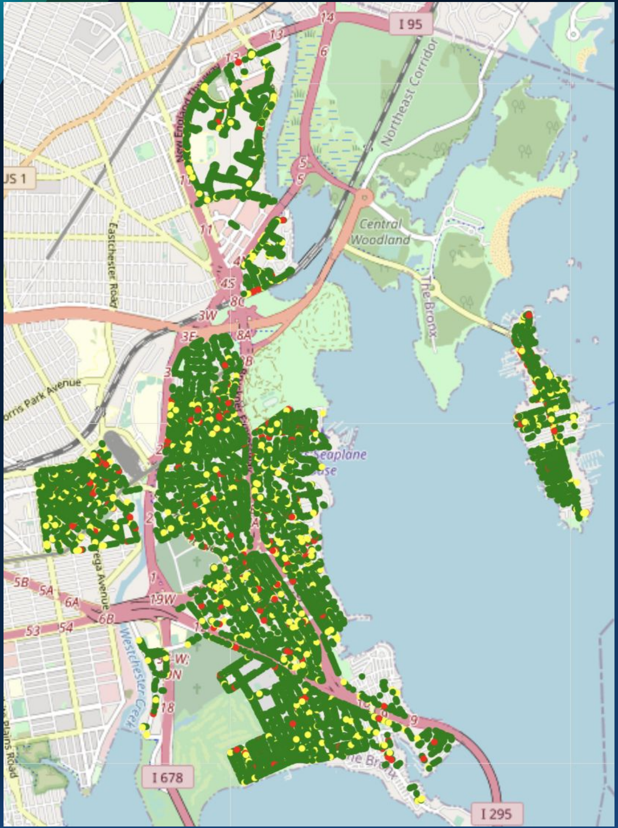
Tuned forest confusion matrix

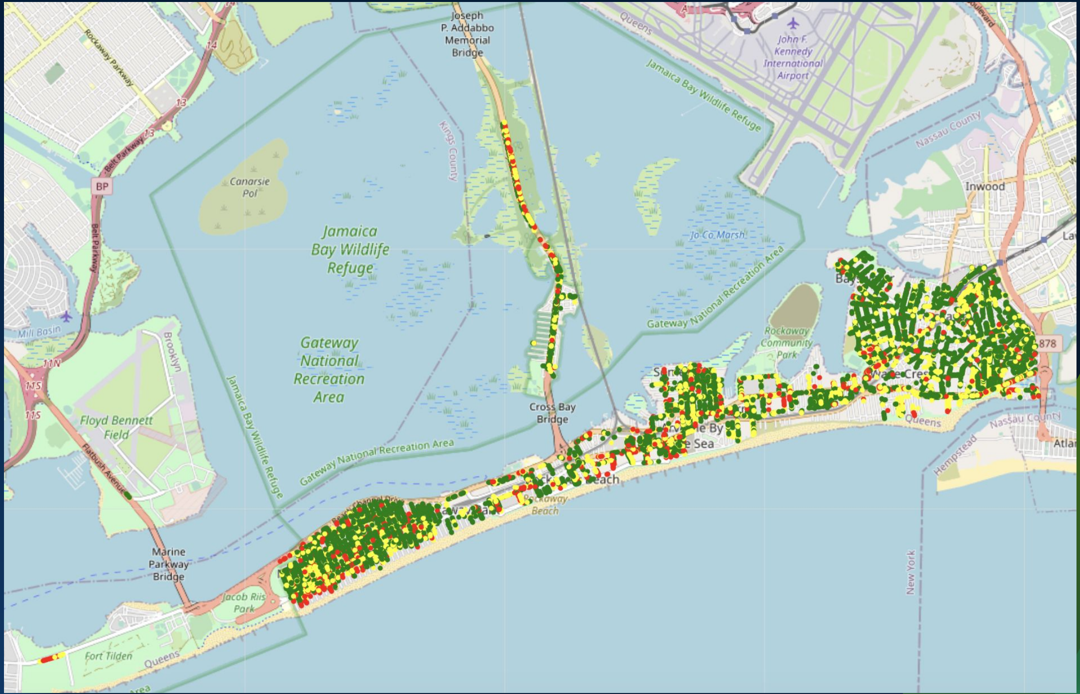
True label \ Predicted label	Fair	Good	Poor
Fair	0.43	0.47	0.10
Good	0.11	0.85	0.04
Poor	0.21	0.35	0.44

Accuracy: **77.1%**
Weighted F1: **78.2%**
Weighted Precision: **77.1%**
Weighted Recall: **79.7%**

Hyperparams: *class_weight='balanced', max_features=11, min_samples_split=11, min_samples_leaf=2, n_estimators=500*

Community Board 210
Neighborhoods: *Co-op City, City Island, Throggs Neck, Country Club, Zerega, Westchester Square, Pelham Bay, Waterbury Lasalle*





Community Board 414
Neighborhoods: *Breezy Point, Belle Harbor, Broad Channel, Neponsit, Arverne, Bayswater, Edgemere, Rockaway Park, Rockaway and Far Rockaway*

Any questions?