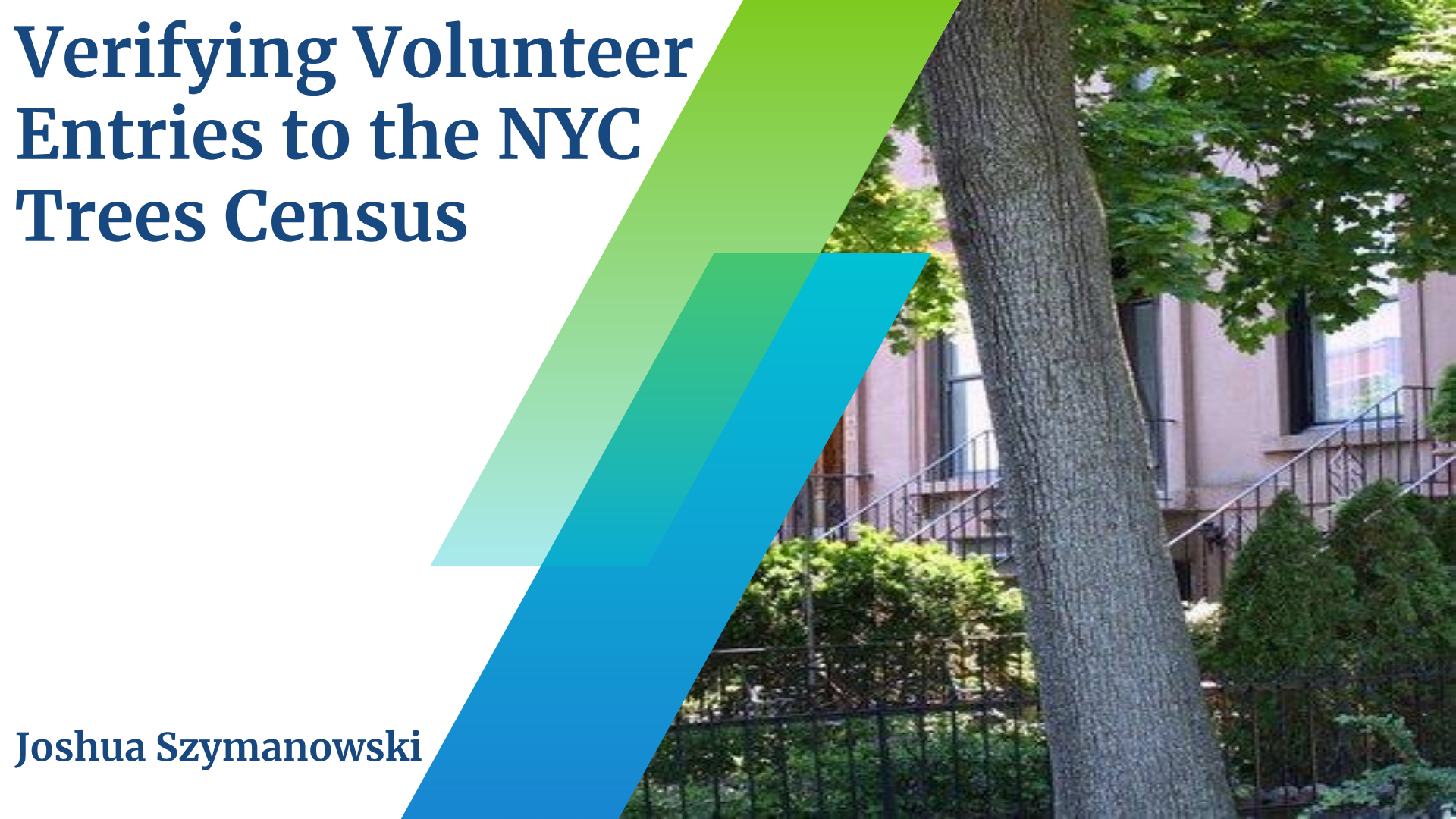
Verifying Volunteer Entries to the NYC Trees Census



Joshua Szymanowski

A look at the data

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 2 3 4 5

Data Cleaning

Drop stumps and dead trees

Drop volunteer data

Turn data objects into numericals

EDA

Effects of location:
- Borough
- Neighborhood
- Political districts

Effect of species

Root, trunk, and branch problems

Tree guards

Tree stewards

Maps

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

Feature Engineering

Distance of nearest tree using GeoPandas

Number of trees on same block

Create dummy variables

Use community board as neighborhood variable

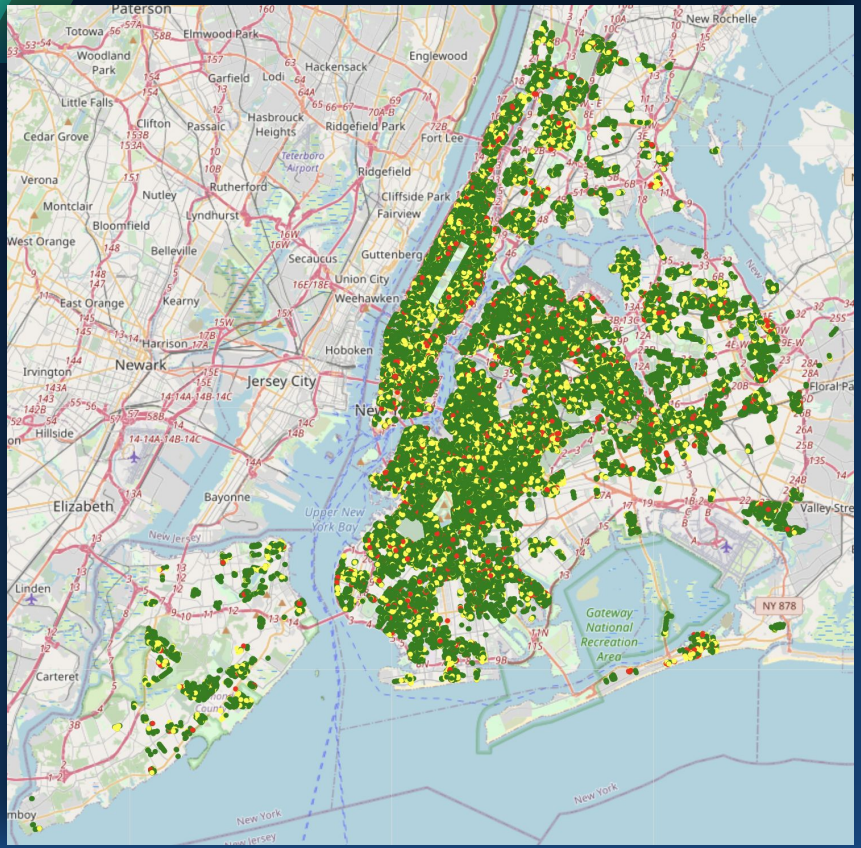
Final Model

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

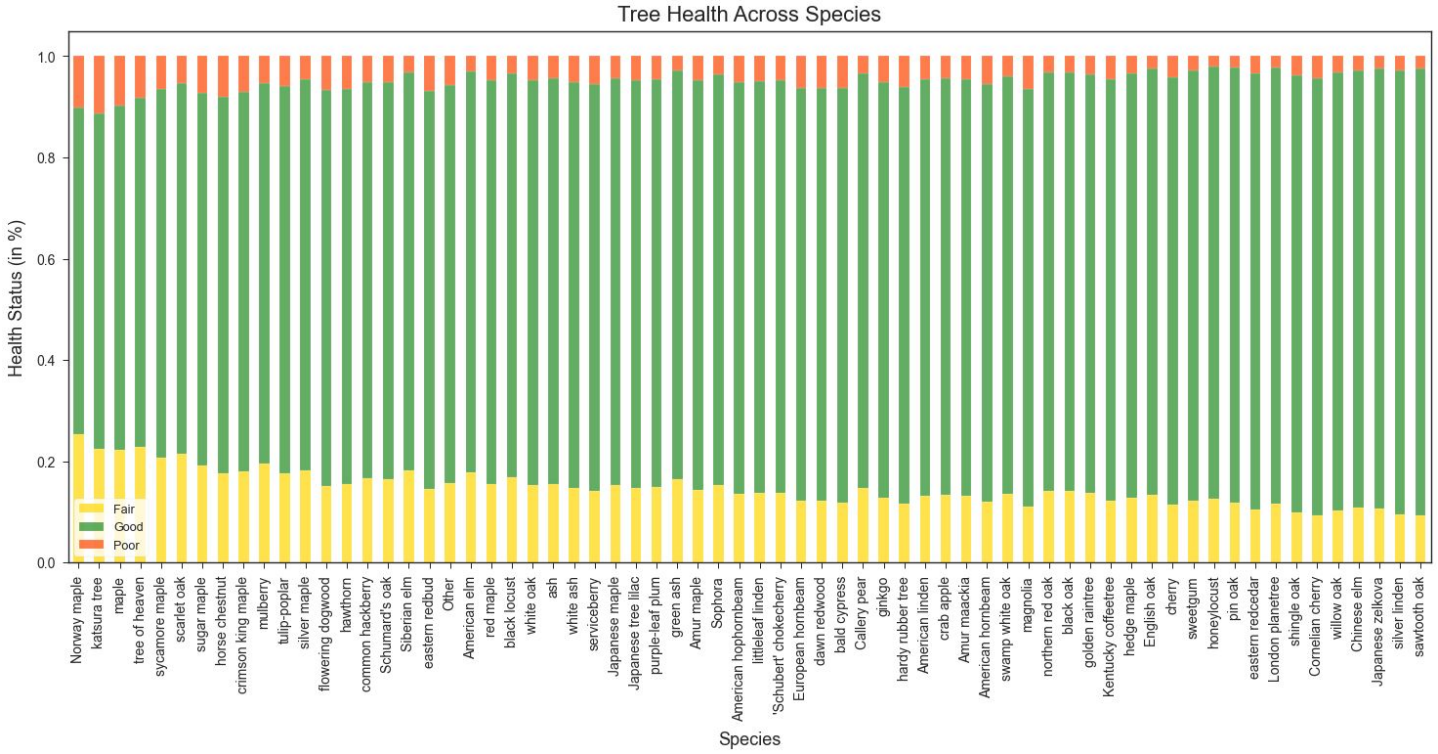
Hyperparams: *class_weight='balanced'*
Accuracy: **84.6%**
Weighted F1: **81.7%**

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



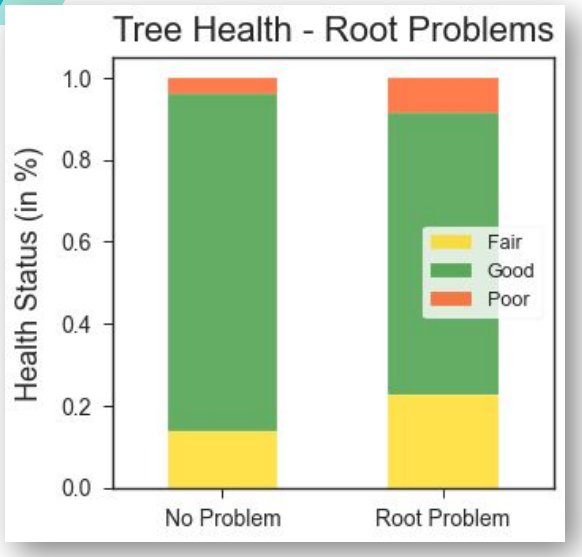
Top 3:

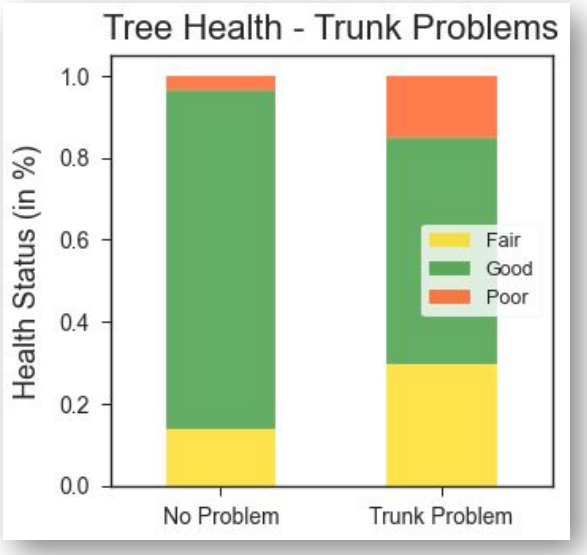
- **Norway maple**
- **Katsura tree**
- **Maple**

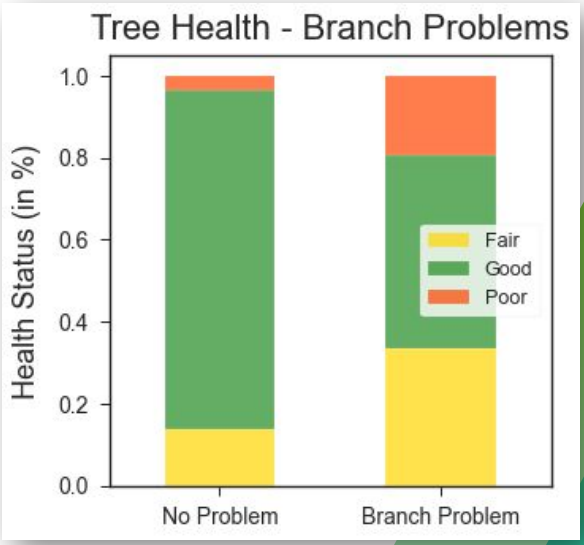
Bottom 3:

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

Tree problems (and solutions)







The most important problems are listed as “Other”.

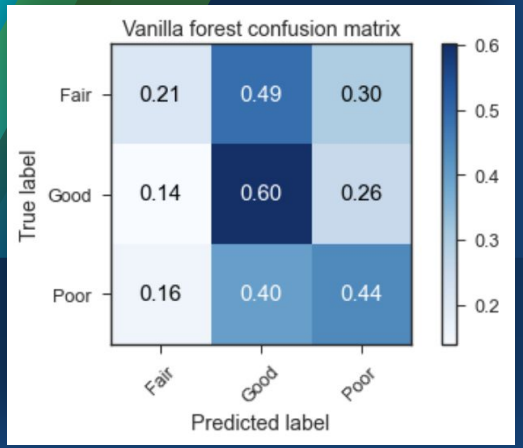
In future censuses, be more specific or have a notes column for each.

Solutions for healthier trees:

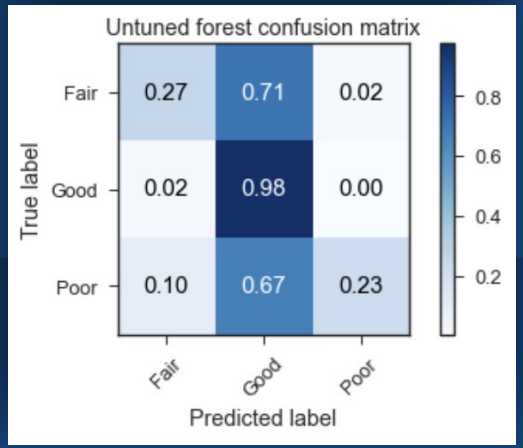
More regular maintenance of trees.

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

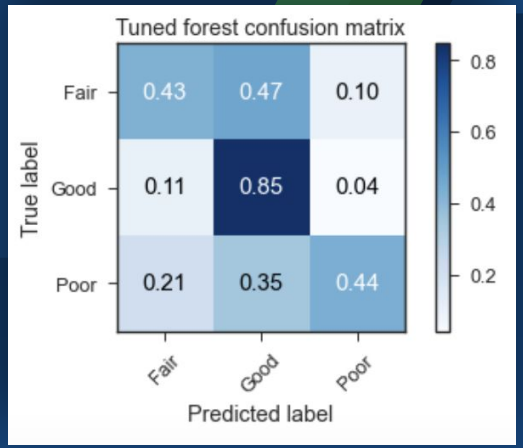
Confusion journey



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



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

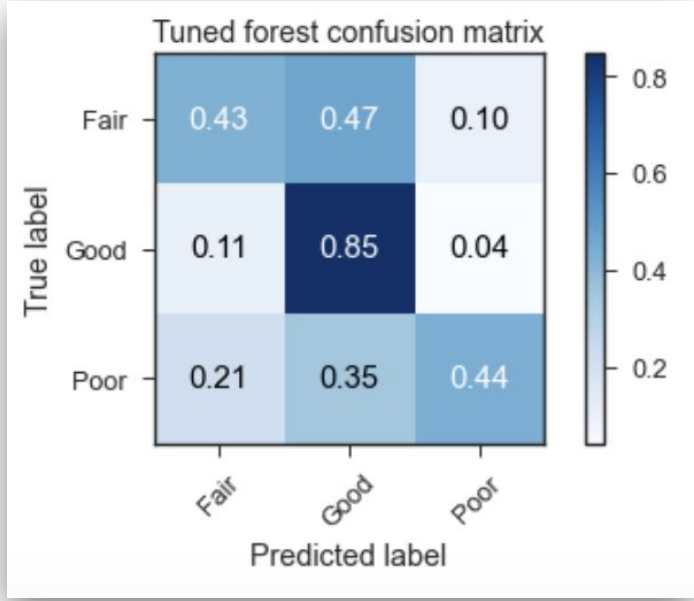


Accuracy: **76.8%**
Weighted F1: **78.0%**

Final model –
Random Forest

Top features (out of 147):

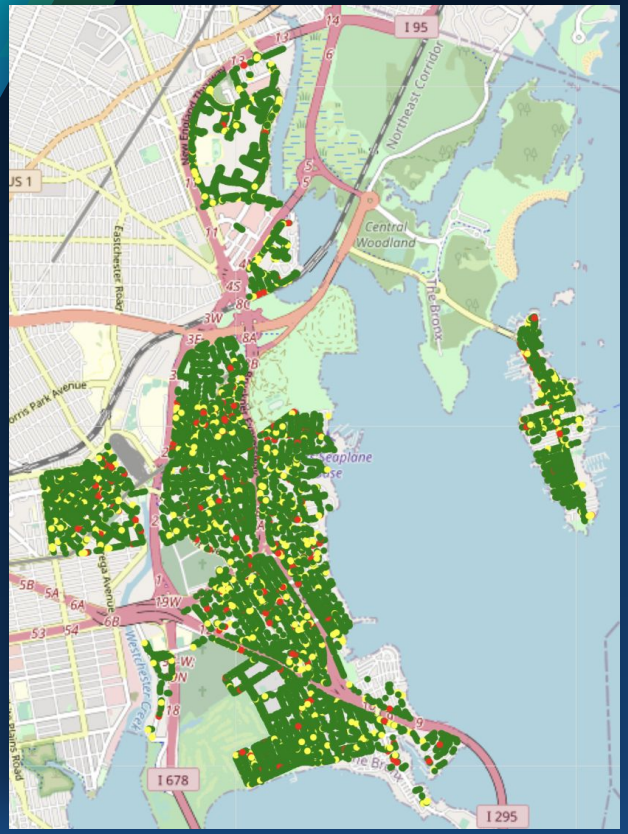
Tree diameter	(14.8%)
Number of trees on block	(11.2%)
Distance to nearest tree	(9.1%)
Sidewalk damage	(7.6%)
Tree stewards [1-2]	(5.1%)
Root problems [stone]	(4.8%)
Species [Norway maple]	(2.6%)
Trunk problems [other]	(2.4%)
Branch problems [light]	(2.2%)
Species [London planetree]	(2.0%)
Branch problems [other]	(2.0%)
Community board [503]	(1.8%)
Species [Honeylocust]	(1.3%)
Borough [Queens]	(1.2%)

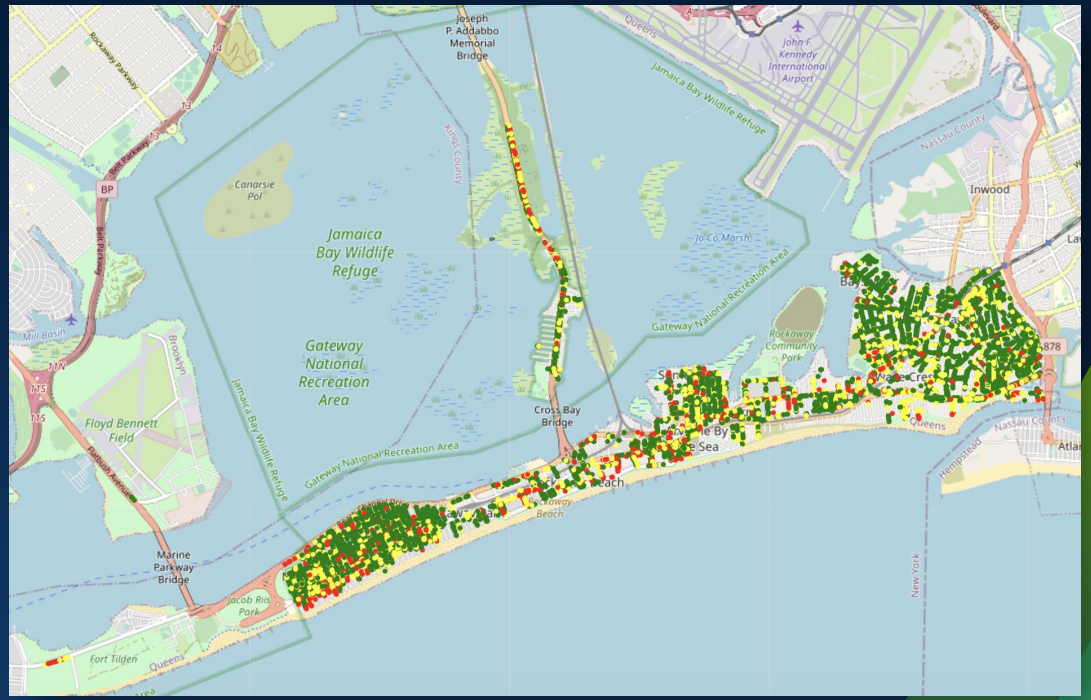


Accuracy: **77.1%**
Weighted F1: **78.2%**
Weighted Precision: **77.1%**
Weighted Recall: **79.7%**

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?