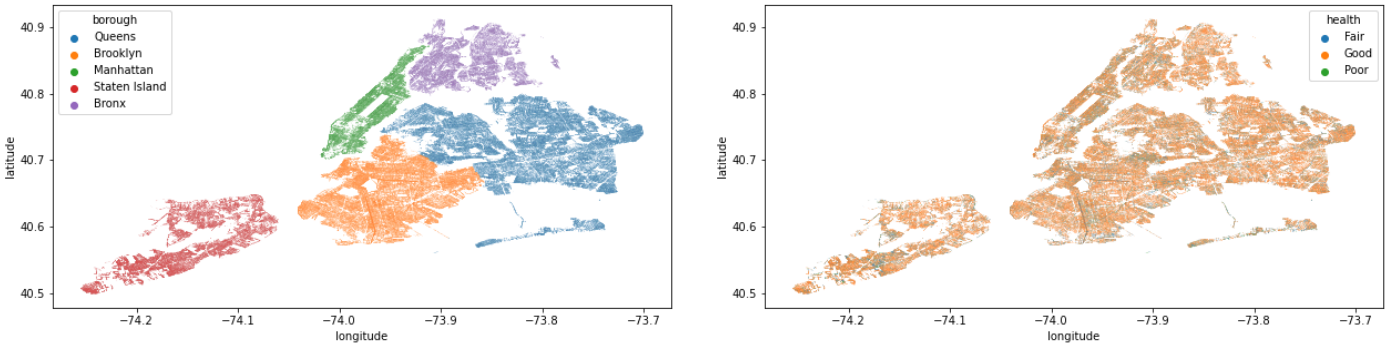
Capstone Two: Final Report

1. Problem Statement:

New York City is the largest city in the United States, and consequently, has one of the largest number of public trees, with a range of health statuses across all regions:



The health of such trees determines the municipal budget for tree maintenance that is necessary for public safety. Thus, the NYC Parks and Recreations department should hope to predict the health status of trees in order to plan for the annual maintenance of trees with the status of 'fair' or 'poor.'

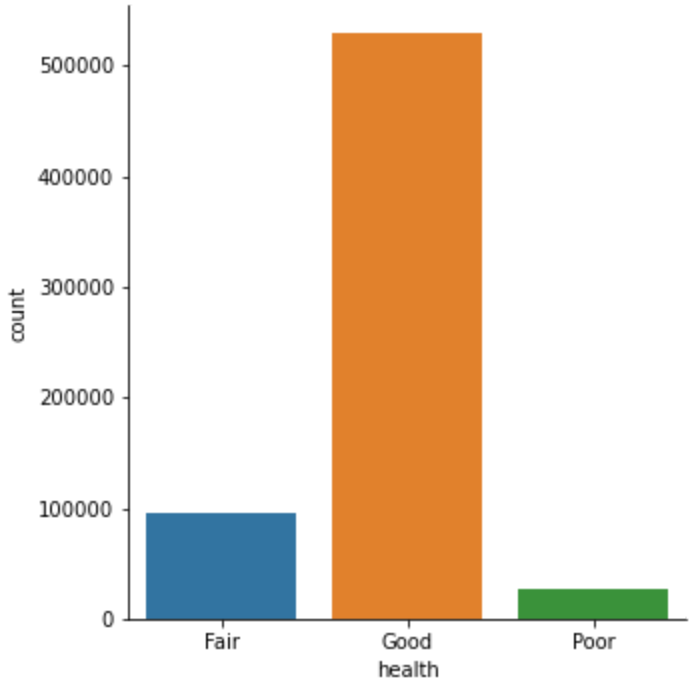
2. Dataset:

In 2015, staff and volunteers partnered with the NYC Parks and Recreations department to collect data on the status of 684k+ trees, as well as notes on 40+ other descriptive features, publicly available here.

3. Data Wrangling/Exploratory Data Analysis:

Target Variable:

There were no duplicate records in the dataset, but there were a number of records with null values for the target variable, the health of trees. Those records were dropped. However, the remaining records with 'health' statuses had very imbalanced classes:



While the fact that the majority of trees are in 'Good' health is positive for the NYC Parks and Recreations department, to address this imbalance, the minority classes were oversampled using SMOTE. Lastly, the target variable was label encoded before use in modeling.

Numeric Predictor Variables:

Excluding latitude and longitude, only two of the predictor variables were numeric. Trees that were in actuality cut-down stumps had null health values (since they were no longer alive). Thus, the variable that measured stump diameter, 'stump_diam', had the value of zero for all records after dropping records with null health values. The stump diameter column was consequently also dropped.

The remaining numeric variable that measured tree base diameter, 'tree_dbh', had several extreme outlier values in the upper range. The records were filtered to exclude the top 5% of tree_dbh values, likely the result of mis-entry by volunteers (*some values were recorded with a diameter of 400+, which is larger than what is likely possible for trees in an urban area like NYC). Lastly, the tree_dbh variable along with latitude and longitude were scaled before use in modeling.

Categorical Predictor Variables:

The remaining predictor variables were categorical. They were all one-hot encoded before use in modeling.

Final Shape:

The final shape of the data used to train the model and test model performance was:
1496298 rows

170 columns
The data was split 80/20 for train/test sets.

4. Modeling:

Three models were attempted: Logistic Regression, Random Forest Classifier, and Decision Tree Classifier.

Logistic Regression:
Initial performance of the model was very poor, with just slightly above 0.5 accuracy in both the train and test sets. Notably, the model had poor performance in recall for the 'fair' health class. The model was rejected early on in the process.

The remaining two models, Decision Tree Classifier and Random Forest Classifier, both had much better initial accuracy scores compared to Logistic Regression, but both were also very overfit to the training data; both had 1.0 scores across the entire classification report. Thus, both models needed hyperparameter tuning to increase performance on the test data.

Random Forest Classifier Hyperparameter Tuning:
RandomizedSearchCV was performed to address the initial Random Forest Classifier overfitting with 3 cross-folds and differing values for the following parameters: n_estimators, max_depth, and min_samples_leaf. After re-running the Random Forest model using the best parameters from RandomizedSearchCV, the model was significantly less overfit, but the accuracy (around 0.58 for both train and test sets) could still be improved.

GridSearchCV was consequently performed with differing values around the range of the best parameters from the RandomizedSearchCV. The model slightly improved but accuracy was still not strong (around 0.6). This model was also rejected.

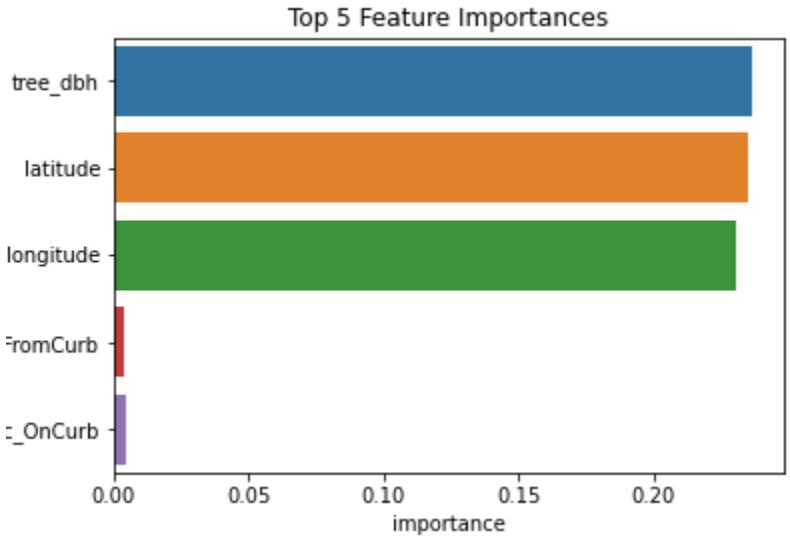
Decision Tree Classifier Hyperparameter Tuning:
Both RandomizedSearchCV and GridSearchCV (similar to above) were again performed with 3 cross-folds for the parameters: max_depth, min_samples_leaf.

Even after pruning, the Decision Tree Classifier seemed to be overfit to the training data, but the overfitting was less after tuning than before, and performance on the test data improved (around 0.75).

Final Model:
The final model was a Decision Tree Classifier with the following parameters:
max_depth: 120

min_samples_leaf: 2

The five features with the most importance to the model were the following:



Of note, the only three numeric predictor variables (tree_dbh, latitude, and longitude) were significantly more important to the model than all other categorical predictor variables. Following, whether or not a tree was on the curb held importance.

5. Application:

The NYC Parks and Recreations department can use the Decision Tree Classifier model to predict the future health of public trees. This will allow the department to strategically allocate resources toward the maintenance of the minority of trees that have the status of 'fair' and 'poor' health. Additionally, by analyzing what features contribute to the 'good' health of trees, the department can invest resources for future plantings.

6. Future Research:

Although the model was trained on a fairly large dataset, the weighting of latitude and longitude in importance in the final model indicates that the health of trees is very dependent on geographical location. This may imply that more localized analysis, perhaps at the borough level, will be more effective at making predictions. Additionally, this potentially indicates that while the model may be helpful for the city of New York to plan its municipal budget, other cities with different terrain and climates, will not be able to adopt the same dataset for model training. Future research by other cities can include the recording of such data.