Save New York City Trees

Introduction:
New York City conducted a street tree census in 2015-2016 with the help of more than 2,200 volunteers. This was the largest participatory municipal urban forestry project in United States history. Using both high tech tools and survey wheels, tape measures, and tree identification keys, citizen mappers helped create a spatially accurate digital inventory of NYC's Street trees. This included:

- **Number of Trees.** Surveyors mapped 666,134 street trees on 131,488 blocks in New York City, walking a total of 11,093 miles.
- **Number of Volunteers**. The 2,241 volunteers is double the number that participated in 2006. Volunteers completed 34 percent of the census.

Objective:
The objective of this project was to predict the health of the tree based on various parameters available in the dataset. Ability to predict the health of the tree will be beneficial for Department of Parks and Recreation as they work to address the causes of poor health and avoid plant reaching that stage.

Dataset:
The data can be accessed by New York City's open data website. The data comprises of 683,788 rows and 4 columns. Each row represented a tree in NYC.
https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh

Approach Overview

- Data Wrangling
- Exploratory Data Analysis
- Statistical Analysis
- Modeling
    - Dealing with Imbalanced Data
    - Hyperparameter Tuning

Step 1: Data Wrangling

The main purpose of this step was to make data useable. First step in the process was to check the quality of the data – checked the data for any duplicates and missing values. There were 31619 rows with missing data for health column. Most of these 31619 rows were for dead and stump of trees and 9559 rows were missing location data. Therefore, all the rows with any null values were dropped from the data set leaving behind 642961 rows with alive trees.

Another important step considered for quality check of data was data type. All columns except *created_at* were in expected data type. The *created_at* column was converted to datetime format.

Step 2: Exploratory Data Analysis

The cleaned data was used to perform exploratory data analysis, used graphics and visualizations to explore and analyze a data set.

Below are the main characteristics of the data set:

Tree count by Health

Majority of trees (81.14%) are in good health. Fair (14.77%) and poor (4.10%) health of tree are minority. This implied before applying machine learning models the imbalance in the data needs to be addressed.

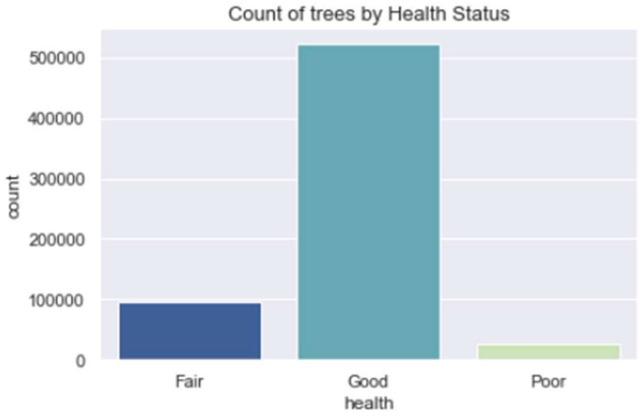
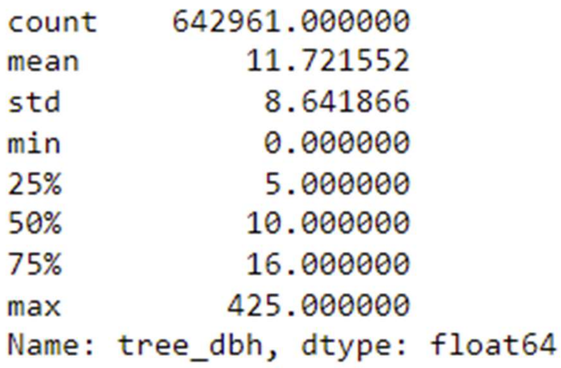


Fig : 1: Count of Trees by health

Location of trees

Most the trees were planted on the curb side. Only 4% of trees were planted off the curb side with only 912 of those were in poor health.

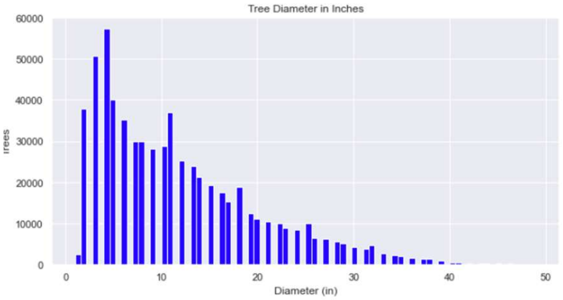
Diameter of the trees



```
count    642961.000000
mean         11.721552
std           8.641866
min           0.000000
25%           5.000000
50%          10.000000
75%          16.000000
max         425.000000
Name: tree_dbh, dtype: float64
```

The diameter of trees had a wide range from 0.0 to 425 inches. The dataset had 220 trees with zero diameter. Technically, its impossible to have a tree with zero diameter. Therefore, 220 of these were dropped from the data set. The number of trees remaining in dataset were 642,741.

Only 353 trees had a diameter of more than 50 inches.



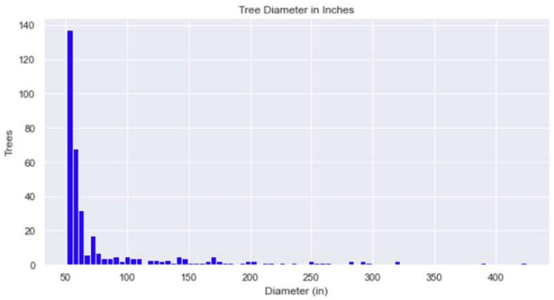


Fig 2: Tree distribution diameter less than 50

Fig 3: Tree distribution diameter more than 50

Species of tree

There are 132 species of trees planted in New York City neighborhoods. London planetree is most found in the city.

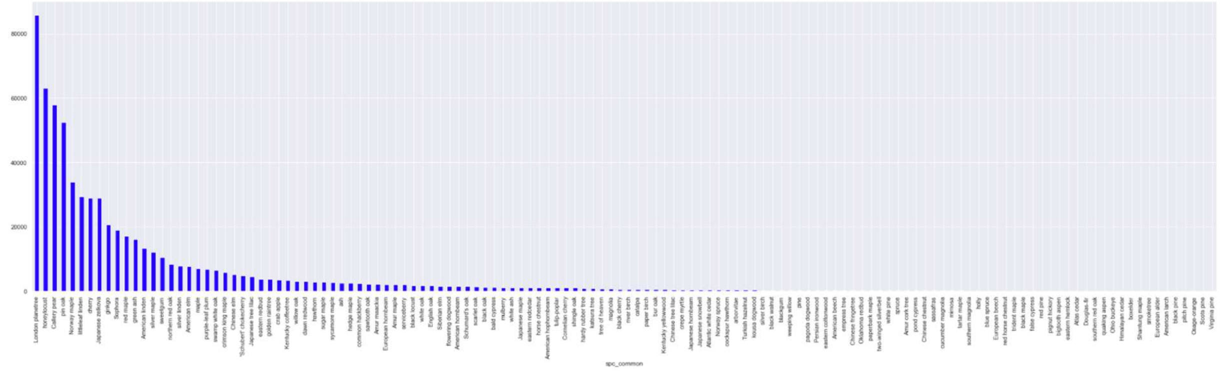
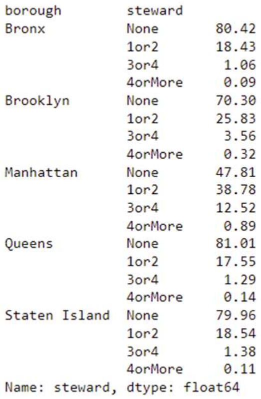


Fig 4: Tree count by species

Stewardship

The Department of Parks and Recreation defines stewardship as the number of unique signs of stewardship observed for a tree. This data point indicates the number of unique signs of stewardship observed for this tree. Not recorded for stumps or.



```
borough        steward
Bronx          None        80.42
               1or2        18.43
               3or4         1.06
               4orMore      0.09
Brooklyn       None        70.30
               1or2        25.83
               3or4         3.56
               4orMore      0.32
Manhattan      None        47.81
               1or2        38.78
               3or4        12.52
               4orMore      0.89
Queens         None        81.01
               1or2        17.55
               3or4         1.29
               4orMore      0.14
Staten Island  None        79.96
               1or2        18.54
               3or4         1.38
               4orMore      0.11
Name: steward, dtype: float64
```

Below is a short list of the most common examples of what counts as one stewardship activity

- Helpful tree guards that do not appear professionally installed
- Mulch or woodchips
- Intentionally planted flowers or other plants
- Signs related to care of the tree or bed, other than those installed by Parks
- Decorations (not including wires or lights added to the tree)
- Seating in the tree bed, usually as part of the tree guard
- Viewing someone performing a stewardship activity during the survey

Guards

Indicates whether a guard is present, and if the user felt it was a helpful or harmful guard. Values Harmful, Helpful, and Unsure all indicate that a tree guard is present. A tree guard is considered 'helpful' if it doesn't impede water getting to the tree and does not raise the soil level or trap debris in the pit.

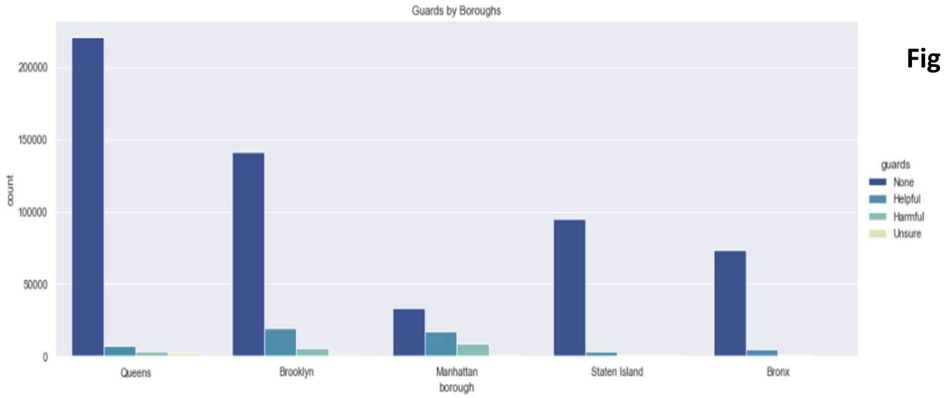


Fig 5: Guards by Borough

User Types

This field describes the category of user who collected this tree point's data. The tree data was collected byTreesCount Staff, Volunteers and NYC Park Staff. More than 40% of tree data was collected by Tree Count staff

Problems

The problems column provided all the problem observed with a particular tree. There were nine types of problems:

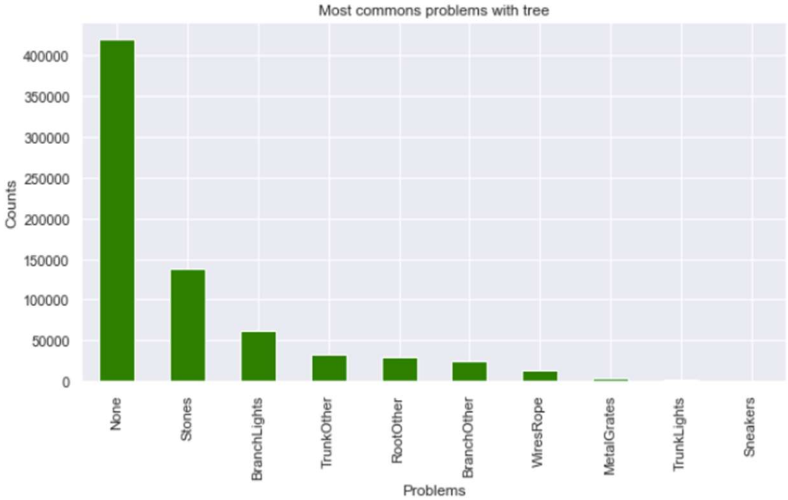


Fig 6: Most common problems

- Trunk problems caused by rope or wires
- Trunk problems caused by lights
- Presence of other trunk problems
- Branch problems caused by lights or wires
- Branch problems caused by shoes
- Presence of other branch problems
- Root problems caused by paving stones in the tree bed
- Root problems caused by metal grates
- Presence of other root problems

Step 3: Statistical Analysis

The Chi-square test, a statistical hypothesis test, was perform to examine independence between various categorical variables and target variable, health of the tree. All of the categorical values had the p-value of above 0.0, so at significance level of 0.05 the null hypothesis that there is no relationship between 'various categorial variables and 'health' was rejected.

For numerical variable – diameter of tree, plotted the distribution of tree diameter based on the health of the tree.

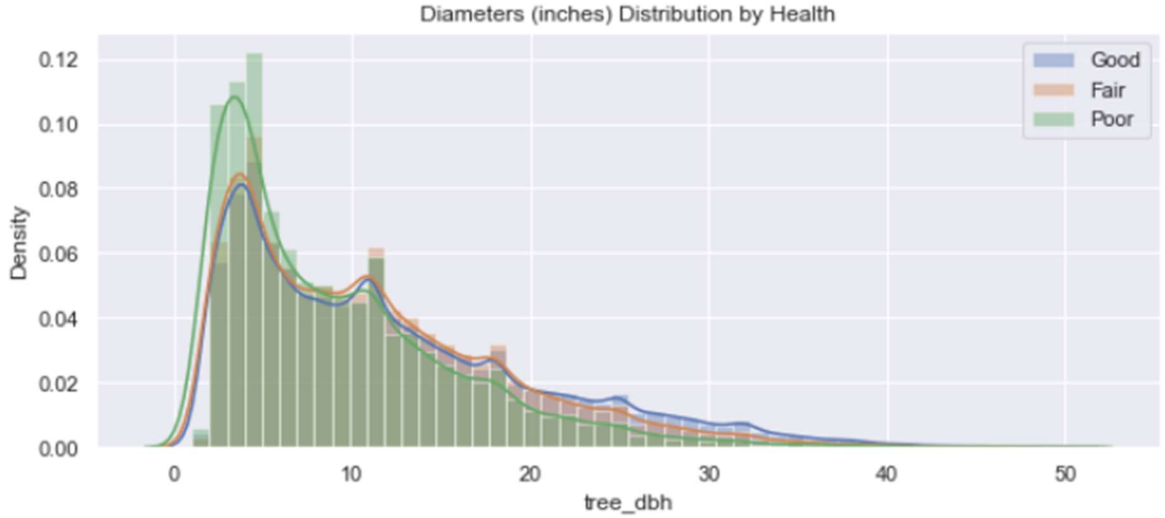


Fig 7: Distribution of diameter of tree by health

Step 4: Modeling

The final step was to build a model to correctly predict the health of a tree. Since this is a multivariate classification problem, along with accuracy score the classification report, which has the precision, recall, and f1-score is also important.

Initial Models

Models	Logistic Regression		K-Nearest Neighbor		Decision Tree		Random Forest		Gradient Boosting	
	Train Accuracy - 0.81		Train Accuracy - 0.85		Train Accuracy - 0.99		Train Accuracy - 0.99		Train Accuracy – 0.81	
	Test Accuracy - 0.81		Test Accuracy - 0.80		Test Accuracy - 0.74		Test Accuracy - 0.81		Test Accuracy – 0.81	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
Good	1.00	0.00	0.94	0.85	0.85	0.86	0.96	0.85	1.00	0.81
Fair	0.00	0.00	0.25	0.40	0.31	0.30	0.20	0.47	0.01	0.51
Poor	0.00	0.00	0.06	0.38	0.18	0.17	0.12	0.39	0.03	0.48

4. 1 Dealing with imbalanced data

Imbalanced classes are a common problem in machine learning classification where there is a disproportionate ratio of observations in each class. In our case, the good health trees were way high in proportion to poor and fair health trees in New York City. Performed oversampling and combination sampling to address the imbalance in data.

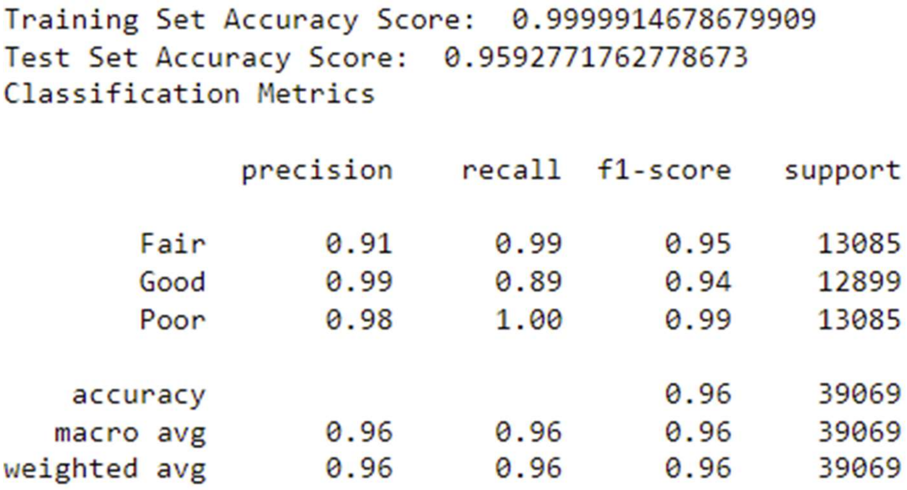
Over sampling methods increases the minority class sample size to match the majority class by duplicating minority samples. The methods used are random over sampler, synthetic minority over-sampling technique (SMOTE), Borderline SMOTE and adaptive synthetic (ADASYN).

Combination under and over sampling methods selectively under sample majority data while oversampling minority data. We use two combination methods, SMOTE-Tomek and SMOTE-ENN. SMOTE-Tomek over samples using SMOTE technique while cleaning with Tomek links. SMOTE-ENN uses SMOTE oversampling and cleans using edited nearest neighbors

4. 2 Hyperparamter tuning

To address the overfitting of data used Cross Validation and Grid search CV for hyperparameter tuning.

Final, model was Random Forest



```
Training Set Accuracy Score:  0.9999914678679909
Test Set Accuracy Score:  0.9592771762778673
Classification Metrics

              precision    recall  f1-score   support

        Fair       0.91      0.99      0.95     13085
        Good       0.99      0.89      0.94     12899
        Poor       0.98      1.00      0.99     13085

    accuracy                           0.96     39069
   macro avg       0.96      0.96      0.96     39069
weighted avg       0.96      0.96      0.96     39069
```

The overfitting of data was not addressed by hyperparameter tuning. Next steps would include to request for more data which is technically not possible.

If the dataset for previous years has the same variables the model can be applied on them and evaluated for its validity on overall dataset for three years – 1995,2005 and 2015.

Lastly, deploying deep learning techniques to address overfitting.