Street Tree Health Prediction Using Machine Learning

NYC 2015 STREET TREE CENSUS

Group 3 –

DATA

SPARTANS

Steve Kofi

Pranav Karthik Chinya Umesha

Mahender Bandi

Girish Sunkadakatte Chandrappa

Executive Summary

- The 2015 NYC Street Tree Census dataset has over 683,788 observations and 45 features was analyzed to understand patterns and factors affecting tree health which is a imbalanced target feature (81.1% Good and 18.9% Poor)
- Our analysis is focused on identifying physical, environmental and species related issues linked to poor tree health conditions with the help of Exploratory Data Analysis and Machine learning modeling.

Key insights from EDA:

- Tree health is imbalanced across boroughs with Manhattan having the highest proportion of poor health trees (24.1%)
- Species like London planetree and Callery pear are abundant compared with other species, show high poor-health rates indicating vulnerable species.
- Root and sidewalk issues are major contributors to poor tree health, especially in Brooklyn (31.9% Root issue and 65.6% Sidewalk issue) and Manhattan (29.1% Root issue and 77.1% Sidewalk issue).
- Stewardship varies by borough with Manhattan showing highest percentage for 3 or 4 stewardship (24.0%) when compared with other.

Data Set Overview

- The dataset collect from the Street Tree Census, organized by **NYC Parks & Recreation with volunteers**.
- Data was collected between May **2015 to October 2016.**
- Dataset comprises of **683,788 observations and 45 feature**s (7 numerical and 38 categorical features).
- Geospatial, Physical Characteristics and Environment characteristics are the important features.
- Target Variable is **health.**

https://www.data-spartans.com/

Category	Features
Tree Characteristics	tree_id, block_id, tree_dbh, stump_diam, spc_latin, spc_common, status, health, problems, steward, guards, sidewalk
Damage Indicators	curb_loc, root_stone, root_grate, root_other, trunk_wire, trnk_light, trnk_other, brch_light, brch_shoe, brch_other
Location & Geography	borough, borocode, boro_ct, nta, nta_name, postcode, zip_city, address, state, latitude, longitude, x_sp, y_sp, , bin, bbl
Civic Boundaries	community board, cncldist, council district, st_assem, st_senate, census tract
Metadata	created_at, user_type

Objectives

- **Understand Tree Health Patterns Across Boroughs**

  Analyze spatial, species wise, and borough wise distribution of tree health using EDA.

- **Identify Key Environmental and Structural Issues Affecting Tree Health**

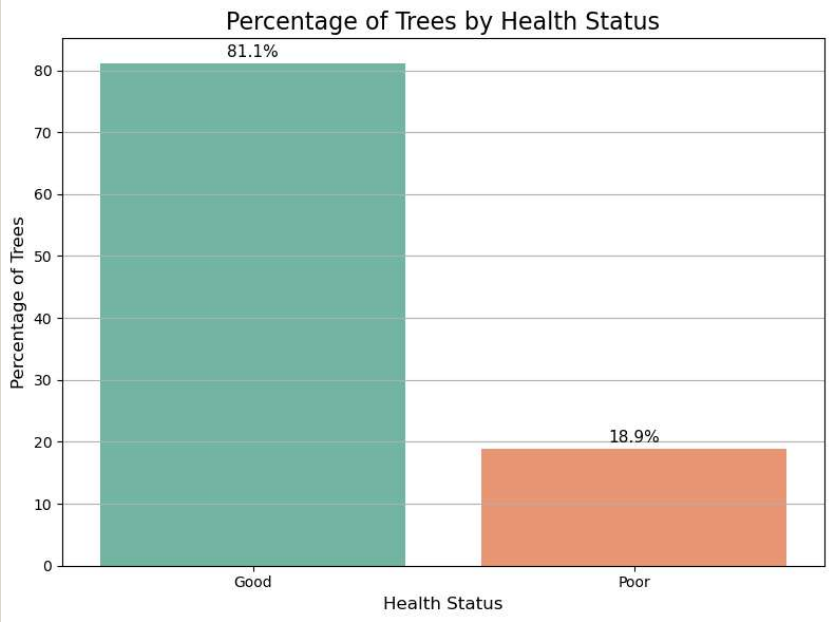
  Analyze root, trunk, branch, and sidewalk problems in relation to poor health cases.

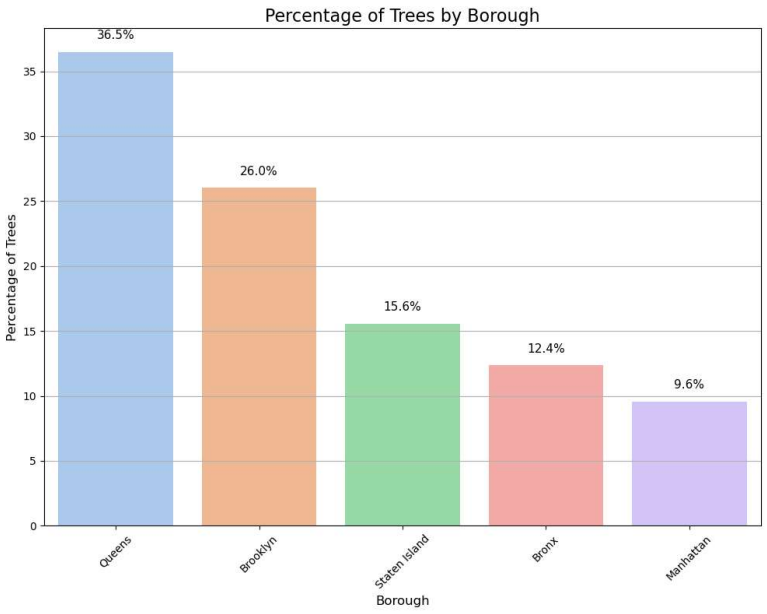
- **Model Tree Health Using Machine Learning**

  Build , Evaluate and Compare models to classify tree health with a focus on minimizing false negatives for poor health trees.

Exploratory Data Analysis

❑ Borough and Target Feature Distribution

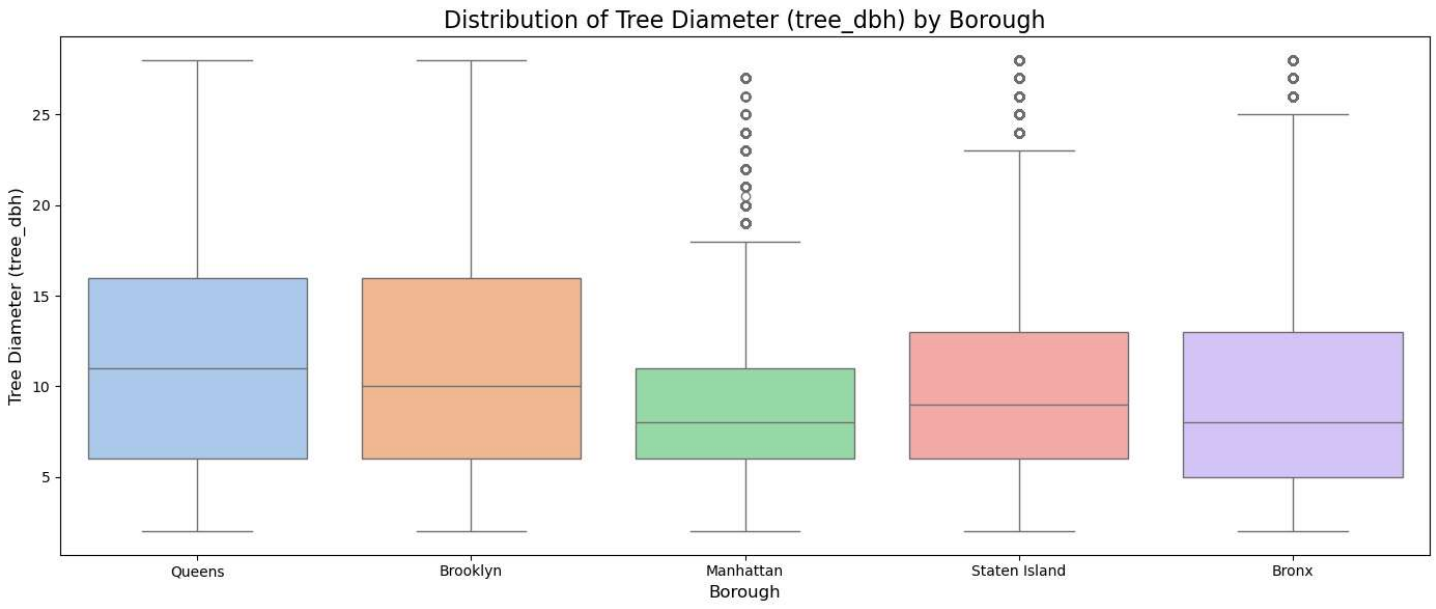




- ❑ Target feature “health” is imbalanced
- ❑ Queen Borough accounts more than one third when compared with other Boroughs.

Exploratory Data Analysis

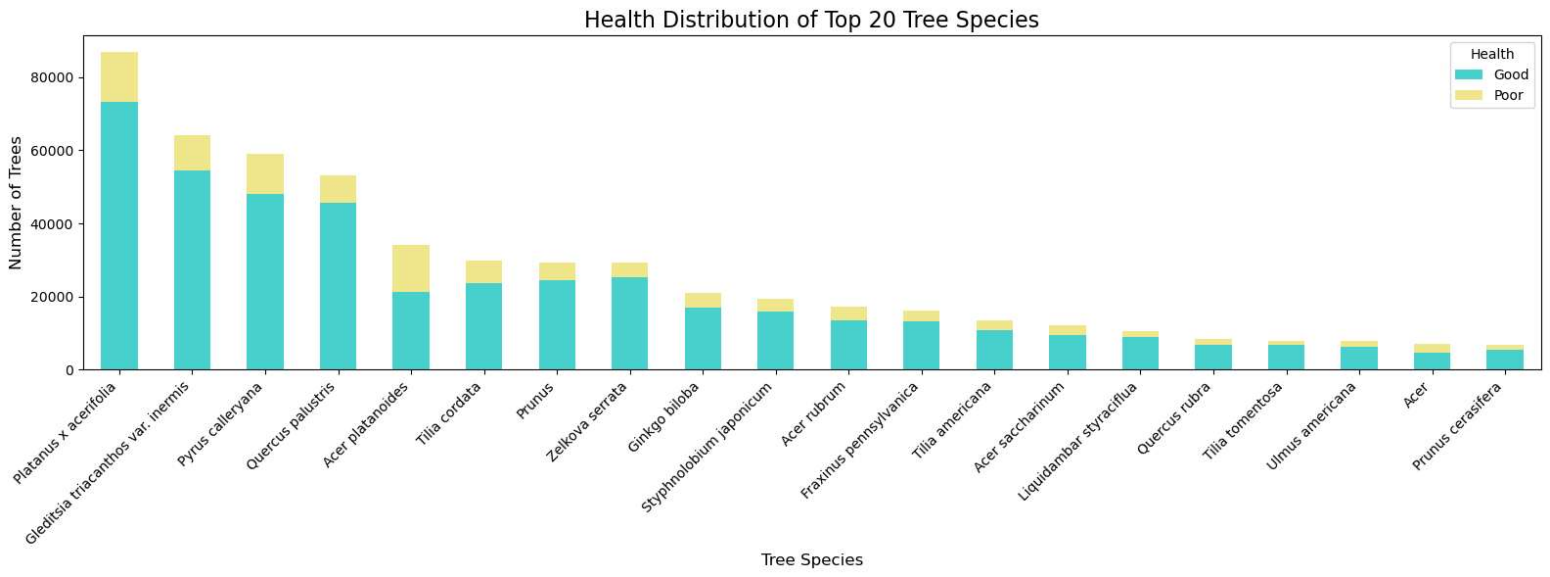
❑ Tree diameter distribution across all Boroughs



- ❑ Queens and Brooklyn Borough have the widest range and larger median tree diameters compared with other Borough's.
- ❑ Manhattan has the smallest median and tighter interquartile range, suggesting younger or more uniformly sized trees.
- ❑ Manhattan and State Island Borough have outliers indicating the presence of very large trees possibly older.
- ❑ Staten Island and Bronx exhibit similar median values, but Bronx shows slightly greater spread in tree sizes.

Exploratory Data Analysis

❑ Health Distribution Among Top 20 Tree Species

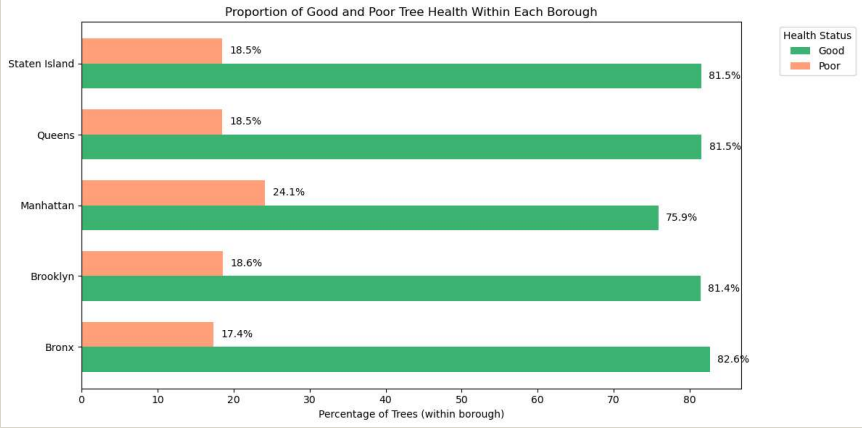


- ❑ Platanus × acerifolia is the most abundant species and has a large portion of trees in poor health. Gleditsia triacanthos var. inermis and Pyrus calleryana also rank high in total count and show noticeable poor health counts.

- ❑ Quercus palustris has high abundance but relatively fewer poor-health trees. Ginkgo biloba and Prunus species have moderate counts, but with a high proportion of trees in good health and small poor-health segments, suggesting they are more resilient and better suited for urban environments.

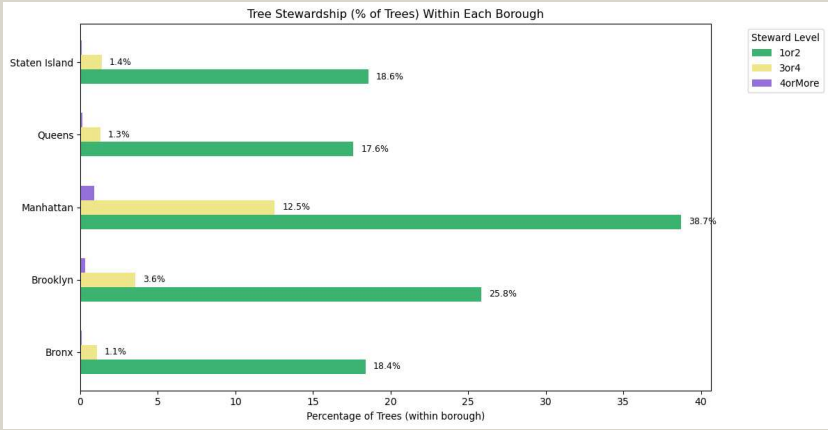
Exploratory Data Analysis

- **What is the proportion of Good vs Poor tree health in each borough?**



- Manhattan has the highest proportion of poor health trees (24.1%), indicating a potential need for improved care .
- Other boroughs like Bronx, Brooklyn, Queens, and Staten Island maintain a better Tree health ratio.

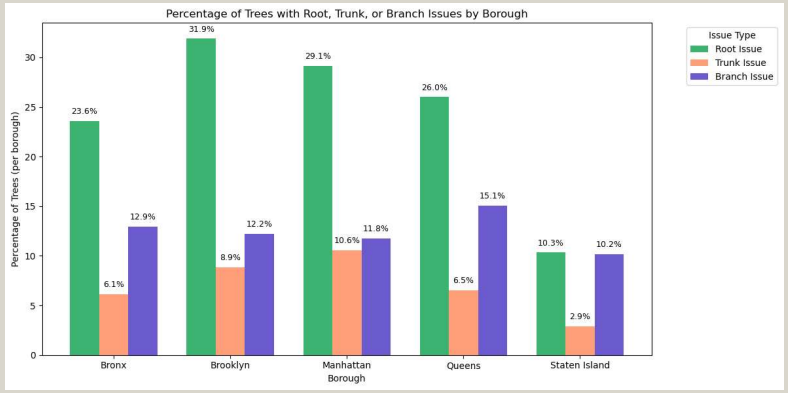
- **How does steward presence vary across boroughs?**

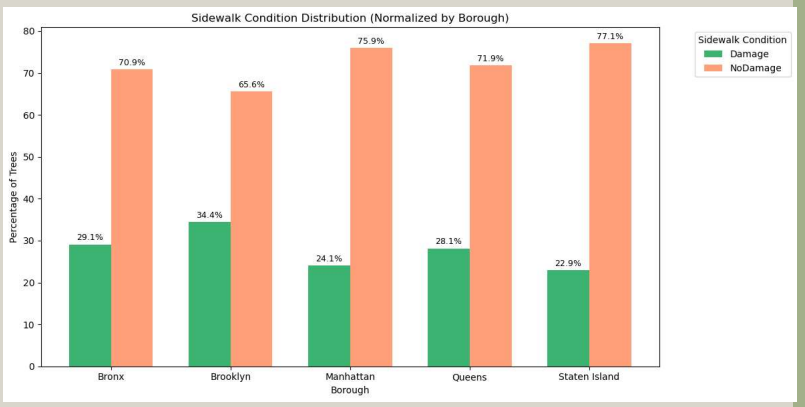


- Manhattan has the highest level of active stewardship, with 12.5% of trees under 3or4 stewardship indicating strong community involvement.
- Bronx, Queens and Staten Island have over 90% of trees with minimal stewardship.

Exploratory Data Analysis

- **How do different types of problems vary by borough?**



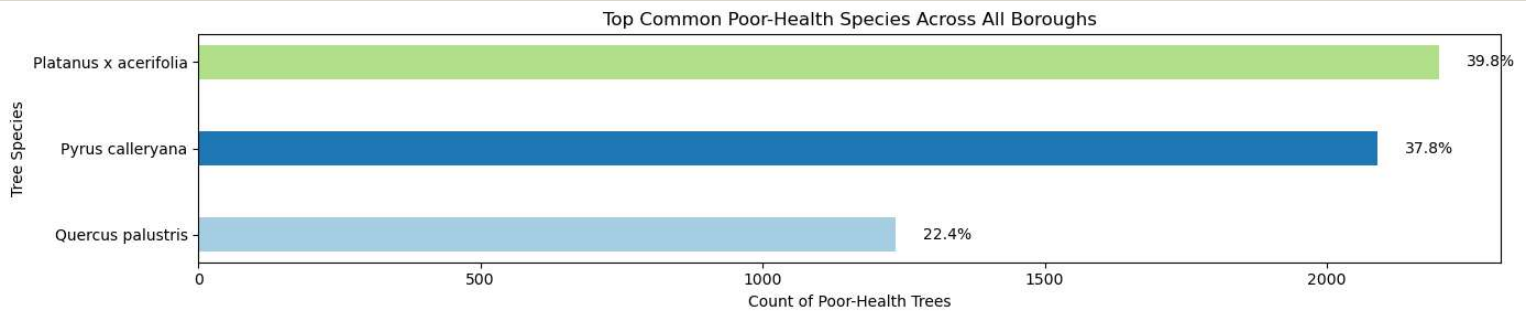


- Brooklyn and Manhattan show the highest percentage of root related issues.
- Staten Island reports the lowest proportion of all tree related issues indication better overall growing conditions.

- Brooklyn has the highest proportion of sidewalk damage potentially caused by tree roots or aging sidewalks.
- Staten Island and Manhattan report the lowest sidewalk damage rates, suggesting better sidewalk conditions or less root interference

Exploratory Data Analysis

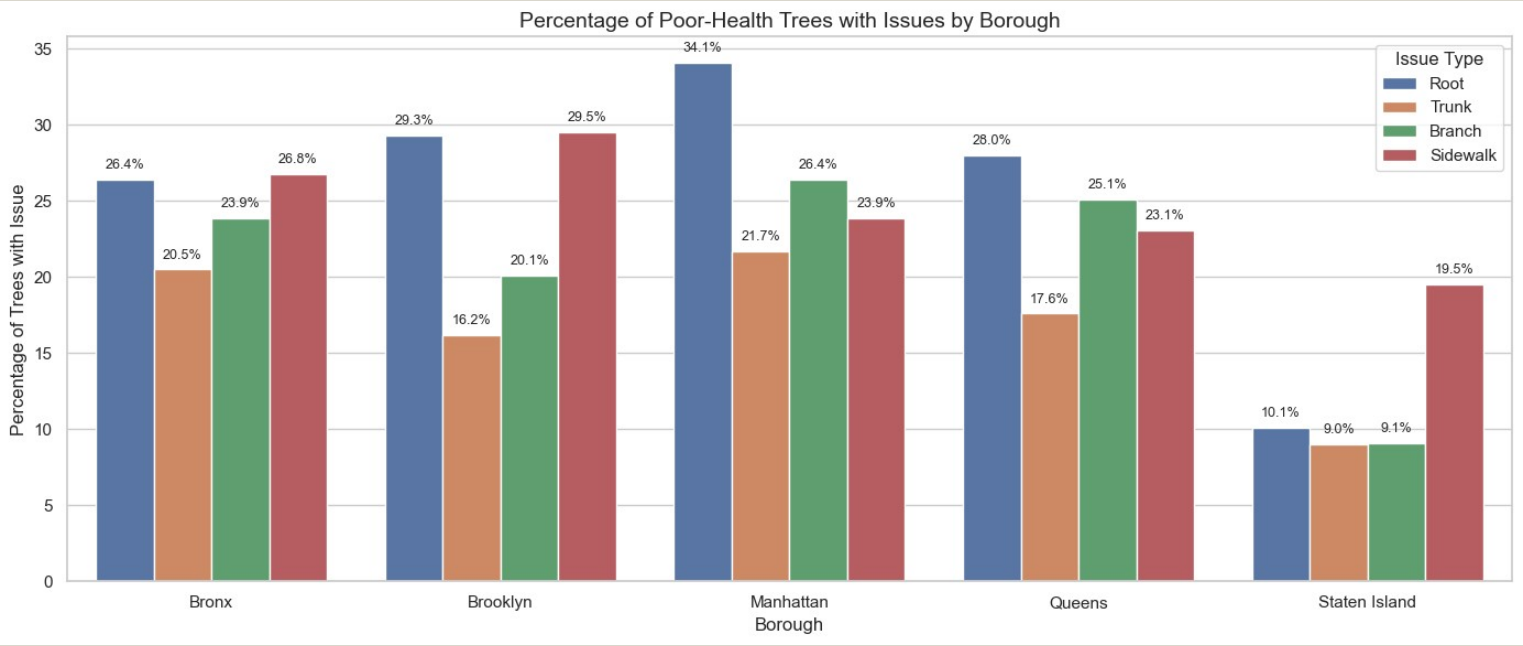
- **Among Boroughs, which ones are most common vulnerable species?**



- **Platanus x acerifolia** is the most frequently found species in poor health across all boroughs indicating a highly vulnerable species despite its popularity.
- Other notable species showing significant vulnerability include **Pyrus calleryana** and **Quercus palustris**, suggesting these species may require closer monitoring.

Exploratory Data Analysis

- **What types of issues are most common in Poor-health trees by Borough?**

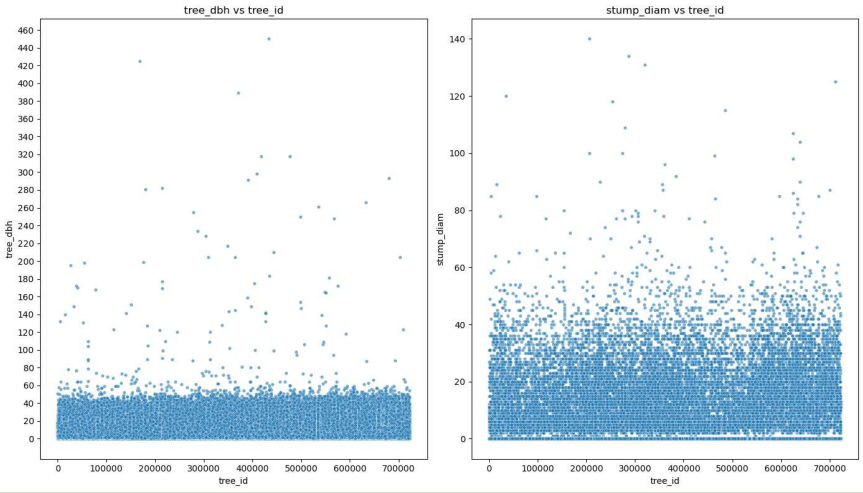


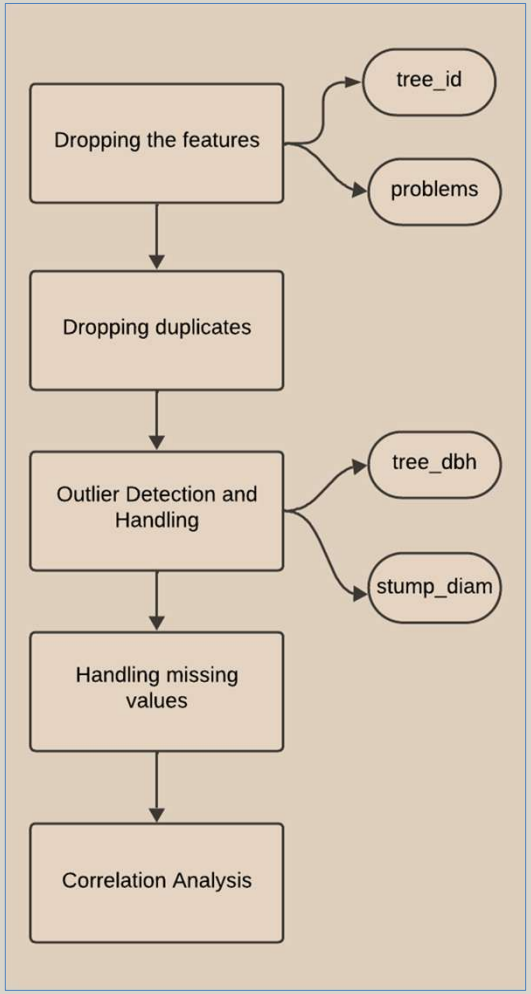
- Root issues are the most common problem in poor-health trees across all boroughs, especially in Manhattan and Brooklyn.
- Sidewalk damage is consistently high among poor-health trees in all boroughs, especially in Brooklyn and Bronx.

Data Preprocessing

- ❑ Outliers in **tree_dbh** and **stump_diam** were identified (IQR Method) and Handled.

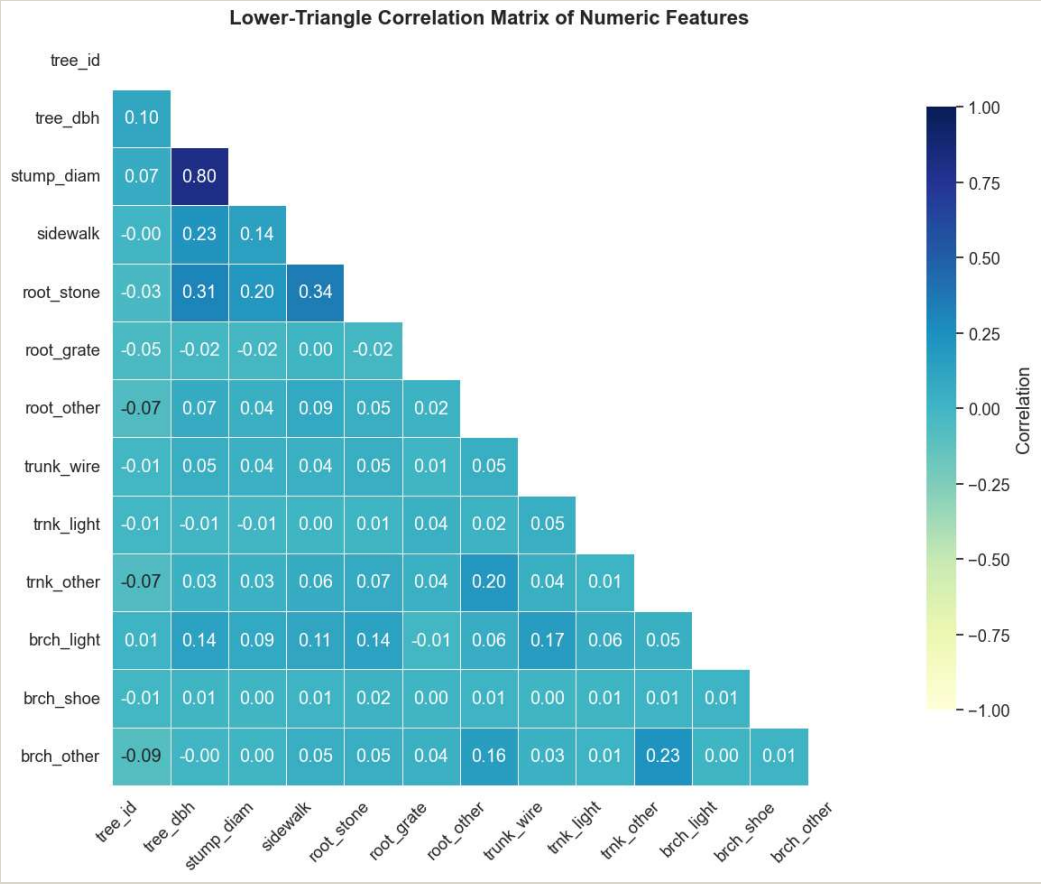
- ❑ Handling Missing values
- ➢ Features with **Tree Status** Stump and Dead were updated as Not applicable.
- ➢ Feature **Steward and spc_latin** are updated with "Not Applicable" and "No observation" respectively.
- ➢ Feature **Sidewalk** is updated with Mode imputation.





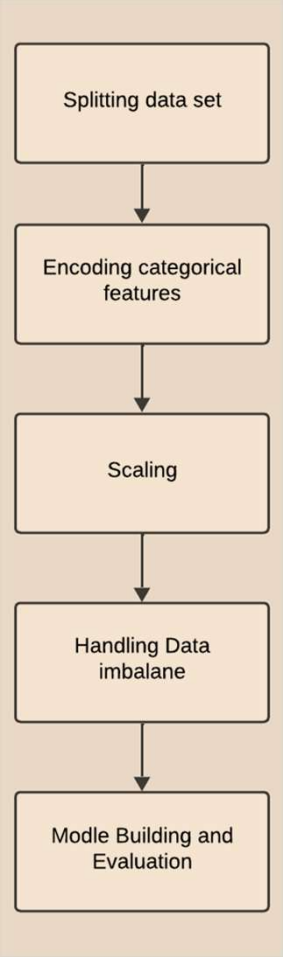
Data Preprocessing

Correlation Analysis



Data Preprocessing

- Target variable was converted into Binary class (Good and Poor) from Multiclass (Good, Fair, Poor).
- One hot encoding for Regression based models and Label Encoding for Tree Based models.
- Min-Max Scaling on the numerical features.
- Under sampling the majority class.



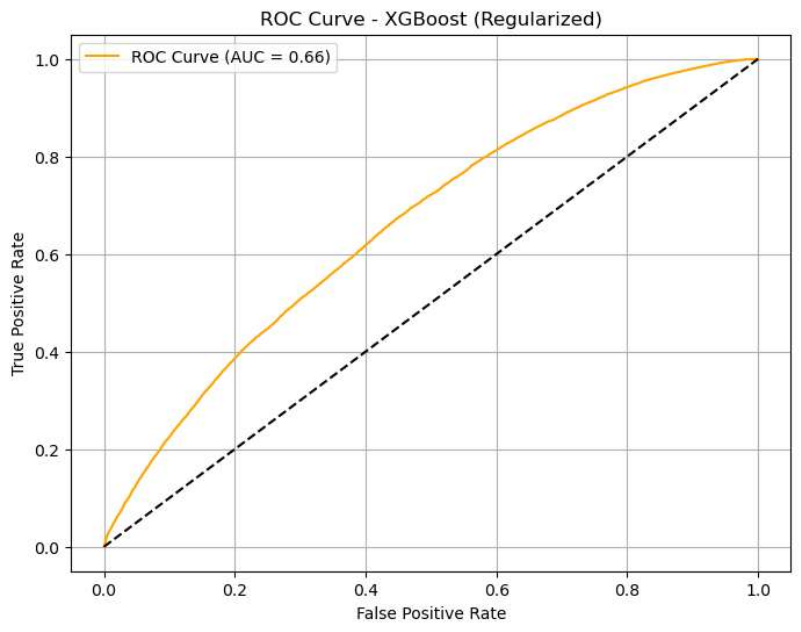
Model Evaluation and Comparison

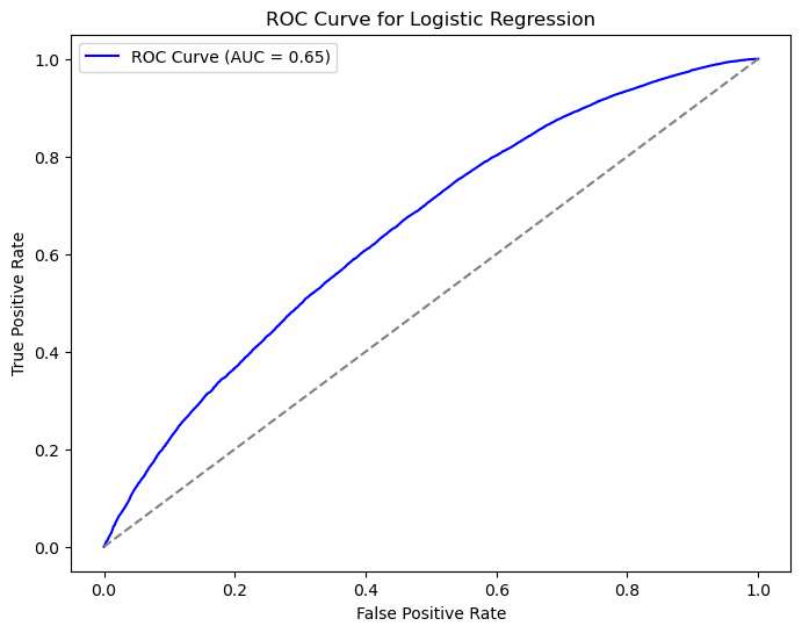
- Both models perform almost similar in identifying Poor health trees (Class 0).
- XGBoost shows slightly better performance across key performance metrics than Logistic Regression Model.
- Logistic Regression is easier to interpret, while XGBoost captures complex patterns more effectively making it better suitable to understand complex patterns.

Model	Accuracy	Precision	Recall	F1 Score	AUC	Recall (Class 0)
Logistic Regression	0.6700	0.7508	0.6700	0.6991	0.6519	0.5006
XGBoost	0.6795	0.7540	0.6795	0.7067	0.6604	0.5010

Model Evaluation and Comparison

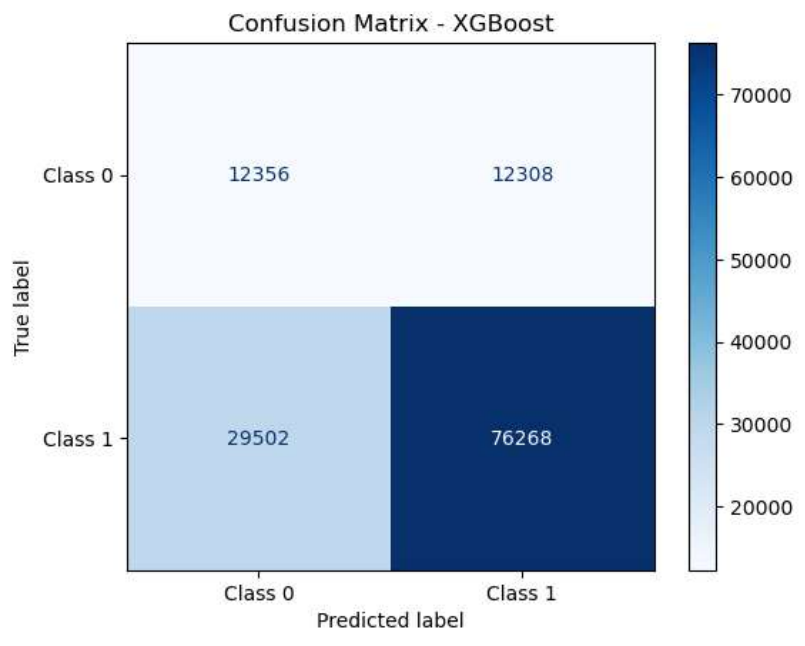
ROC Curve

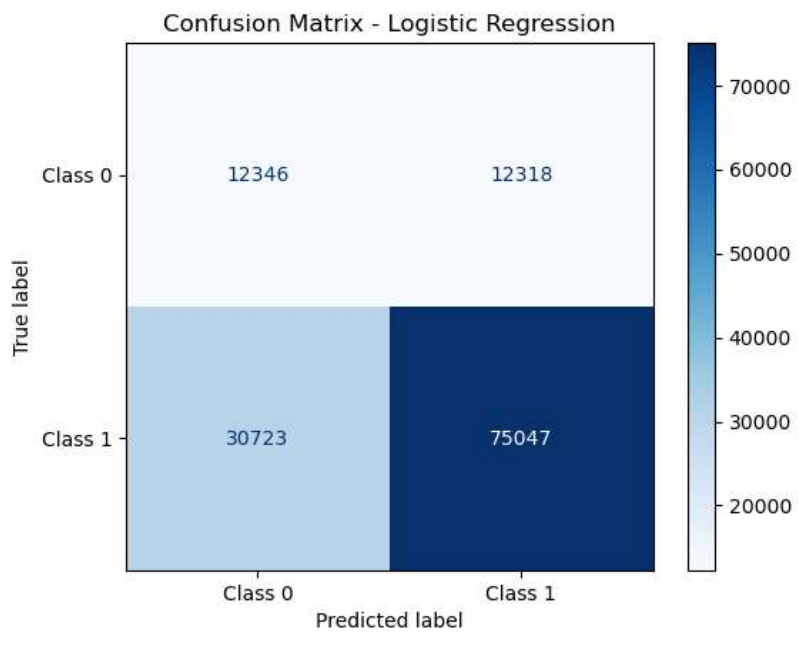




Model Evaluation and Comparison

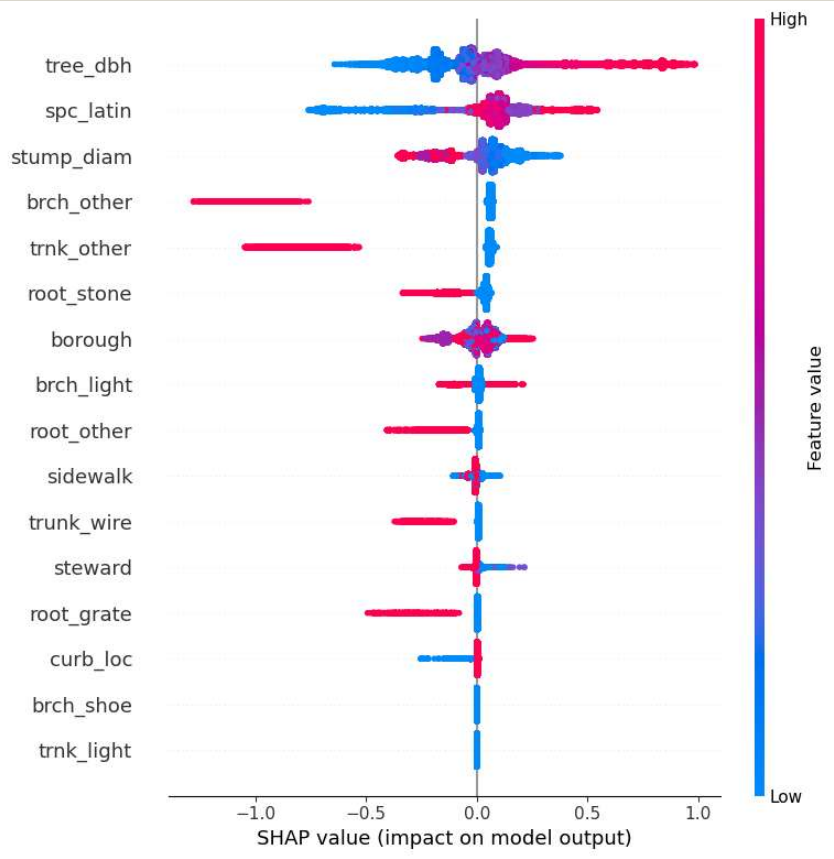
Confusion Matrix





Model Evaluation and Comparison

Feature Importance - XGBoost



Summary

- Both Logistic Regression and XGBoost perform similarly in identifying poor health trees (Class0).
- XGBoost slightly outperforms Logistic Regression in overall metrics such as Precision, F1 Score, and AUC.
- Logistic Regression remains a simpler and interpretable baseline, while XGBoost provides better generalization and complexity handling.

Conclusion

- XGBoost is the preferred model for deployment due to its superior balance between recall for Class 0 and overall performance metrics.
- It effectively captures complex patterns in the data and offers greater reliability for tree health prediction

Future Scope

❑ **Advanced Feature Engineering & Imbalance Handling**

Improve model accuracy by engineering more relevant features and addressing class imbalance with advanced techniques.

❑ **Incorporate Future Tree Census Data**

Use upcoming datasets to improve model generalization and robustness for unseen data.

❑ **Species Trends & Urban Forest Evolution**

Analyze how tree species perform and evolve over time to support biodiversity and long-term urban planning.

❑ **Climate & Socio-Environmental Integration**

Combine tree data with heat island maps, air quality, and socio-economic layers to assess urban greening's role in resilience and equity.

❑ **Interactive Dashboards & Feedback Loop**

Build dynamic monitoring tools and integrate feedback from arborists or the public for ongoing model enhancement.

Thank You.