Tree Troubles

Predicting Sidewalk Damage Resulting From Trees In NYC

Nathan Stevens -- updated: 12/18/2016

credit:http://www.alpinecondoaz.com/

Overview

- Introduction
- Dataset
- Technology Pipeline
- Exploratory Data Analysis
- Unsupervised Machine Learning
- Supervised Machine Learning
- Analysis Application
- Next Steps

Introduction

Tree roots growing under sidewalks often cause cracking or lifting of the pavement once the tree surpasses a certain size. This creates significant tripping hazards for pedestrians, and liability issues for property owners. Furthermore, the cost of repairing such damage is in excess of $100 million per year in the United States[1]. As such, this project seeks to:

1. Predict the likelihood that a particular tree will result in sidewalk damage
2. Elucidate the factors most involved in causing such damage
3. Develop tools to help recommend species of trees and other steps to reduce the likelihood of future sidewalk damage

1. McPherson and Peper 1995; McPherson 2000

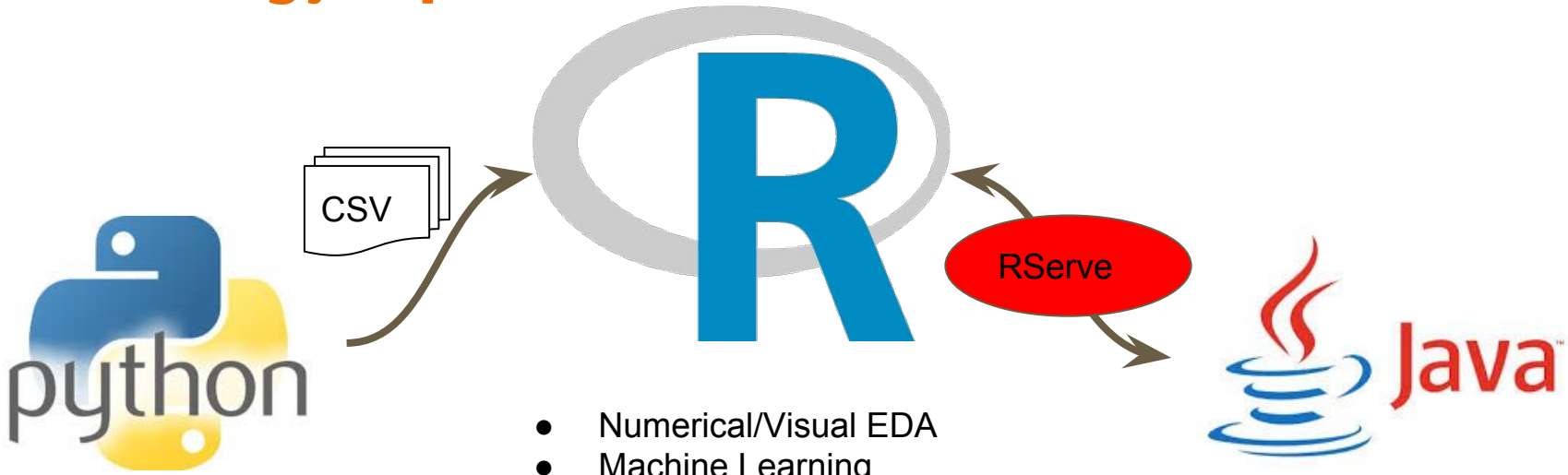
NYC 2015 Tree Census Dataset

In 2015 NYC conducted volunteer-powered campaign to map, count, and care for all of the city's street trees.

- 432,564 Live / 14,099 Dead
- 40% Sidewalk Damage
- Median / Mean DBH[1] (10.0 / 11.6)
- 132 Different Species

DBH -- Diameter at Breast Height / 4.5 ft
photo credit: https://www.nycgovparks.org/trees/treescount/about

Technology Pipeline

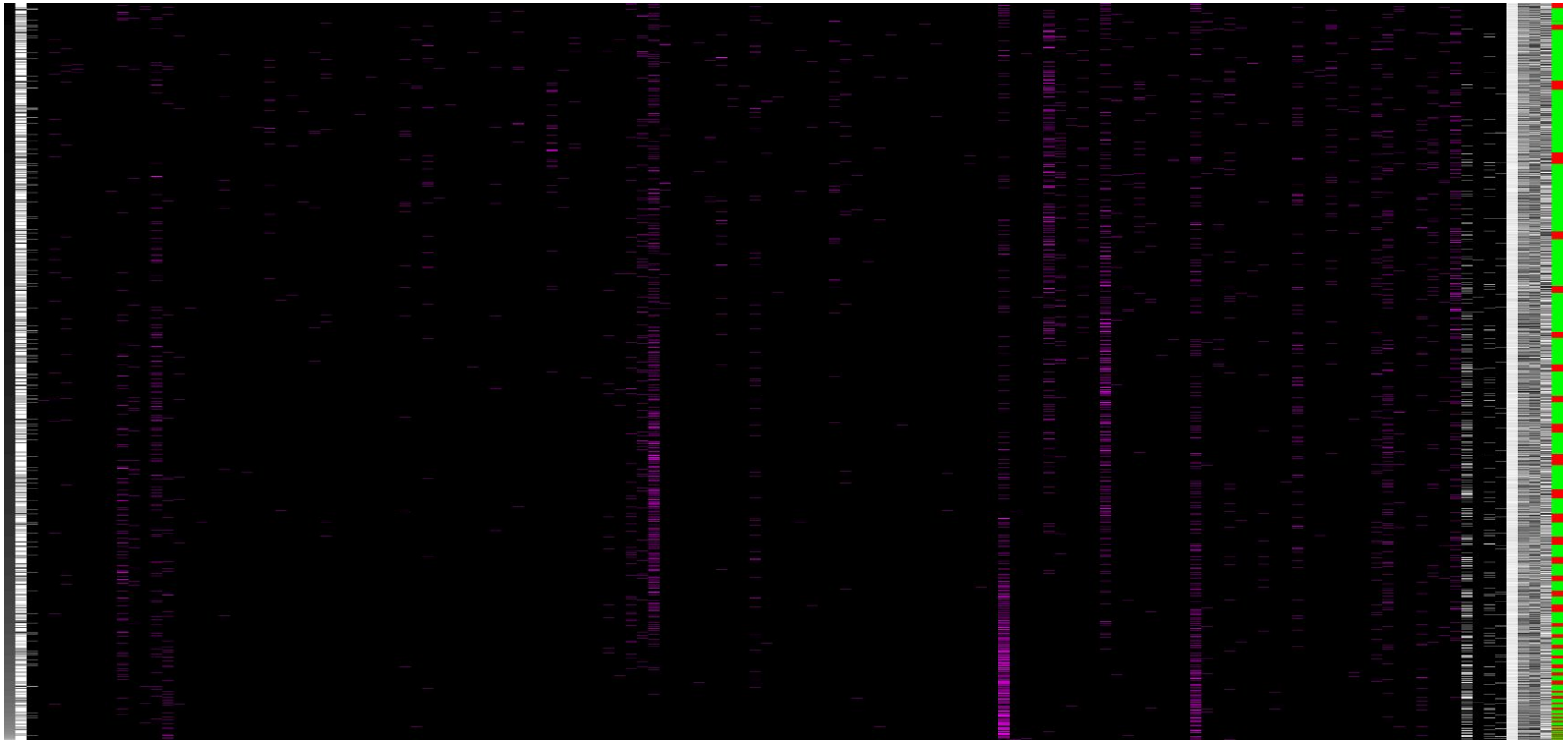


- Data Clean Up
- Data Merging
- Web Scraping

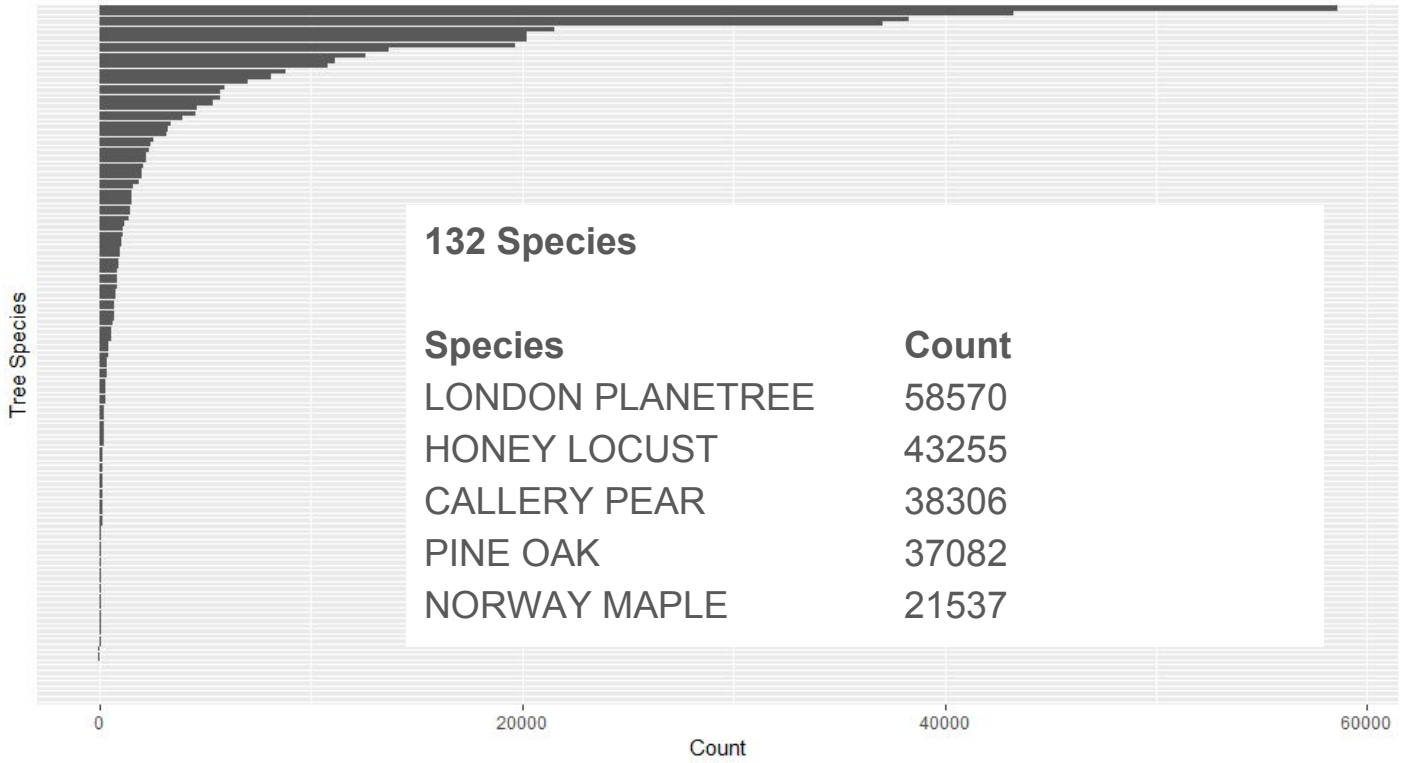
- Numerical/Visual EDA
- Machine Learning

- Visual EDA
- Analysis Application
- Mobile Application

“Heatmap” Data Sorted By Increasing Tree Diameter



Species Barplot



132 Species

Species	Count
LONDON PLANETREE	58570
HONEY LOCUST	43255
CALLERY PEAR	38306
PINE OAK	37082
NORWAY MAPLE	21537

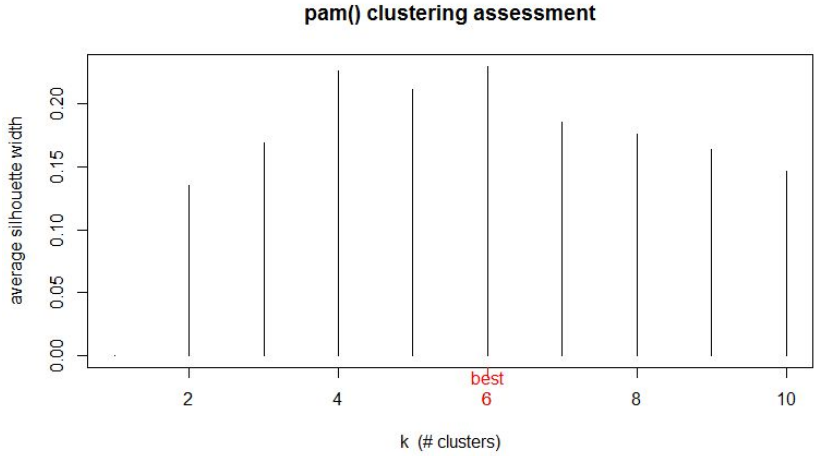
Strength of Association (Cramér's V / ICC*)

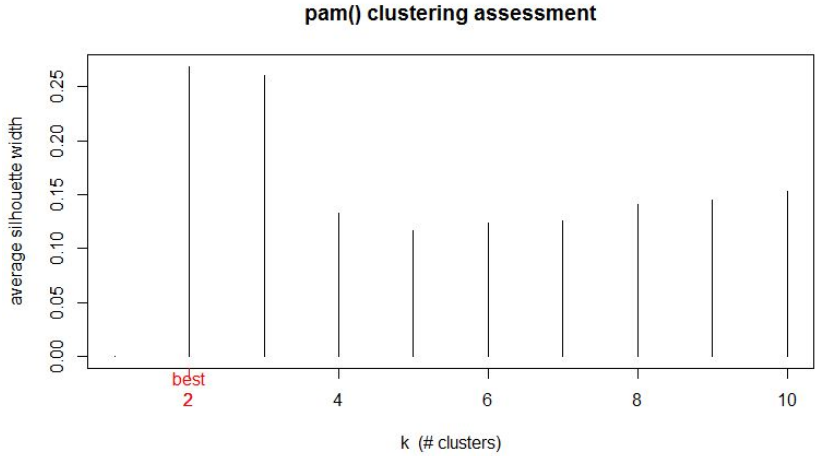
Sidewalk / Tree DBH	0.13*
Sidewalk / Health	0.02
Sidewalk / Species	0.18
Sidewalk / Root Stone	**0.35**
Sidewalk / Root Grate	0.002
Sidewalk / Root Other	0.09
Sidewalk / Trunk Wire	0.03
Sidewalk / Zip Code	0.18
Sidewalk / Boro	0.09

Credit: NYC Park Department

Clustering Results

Clustering was done by first generating a similarity matrix using the “gower” distance then using the “pam” function to find the best number of clusters. Using sample datasets (1000 obs.) which contain geolocation information, the optimal number of clusters is 6. Removing all geolocation related features, with the exception of longitude and latitude, the optimal number of cluster is 2.



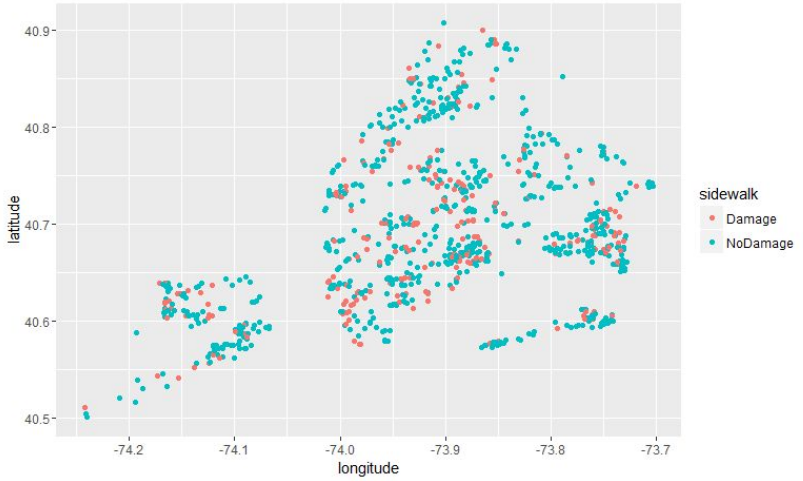


Clustering with Geolocation Features

credit: nycgo.com

Two Clusters vs Sidewalk Condition

1000 obs colored by cluster assignment or sidewalk condition. Almost identical plots with c1 = 263 obs, c2 = 737 obs and “Damage” = 262, “NoDamage” = 738. Chi-squared p-value < 2.2e-16

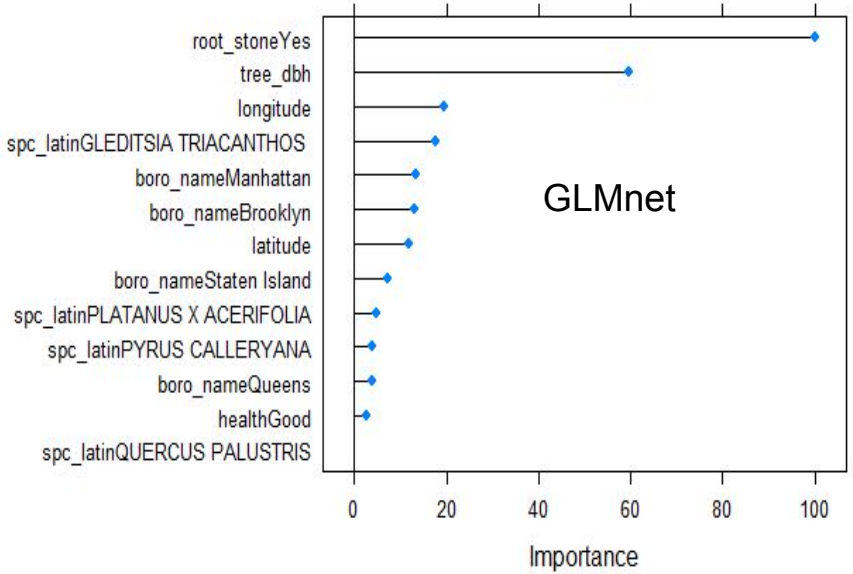


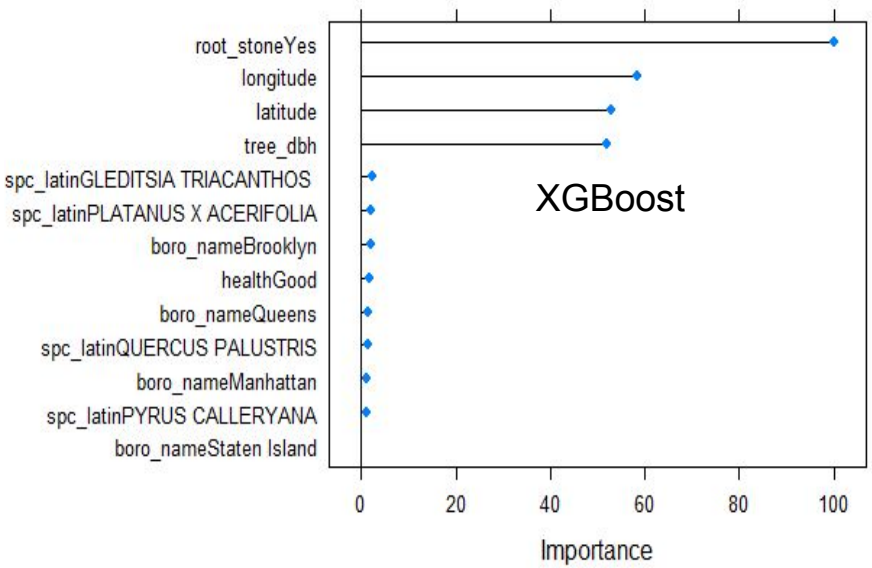
Supervised Learning Results

The Caret package was used to run various machine learning algorithms on full dataset using 80/20 train/test split on an EC2 instance (8 cores/16GB)

- Logistic Regression
  - GLM: 75.8%
  - GLMNet: 75.8%
- KNN: 76.9%
- Naive Bayes: 73.2%
- Tree Based Classification
  - GBM: 77.3%
  - XGBoost: 77.9%
- SVM (radial kernel): 77.1%
- Neural Net: 77.4%

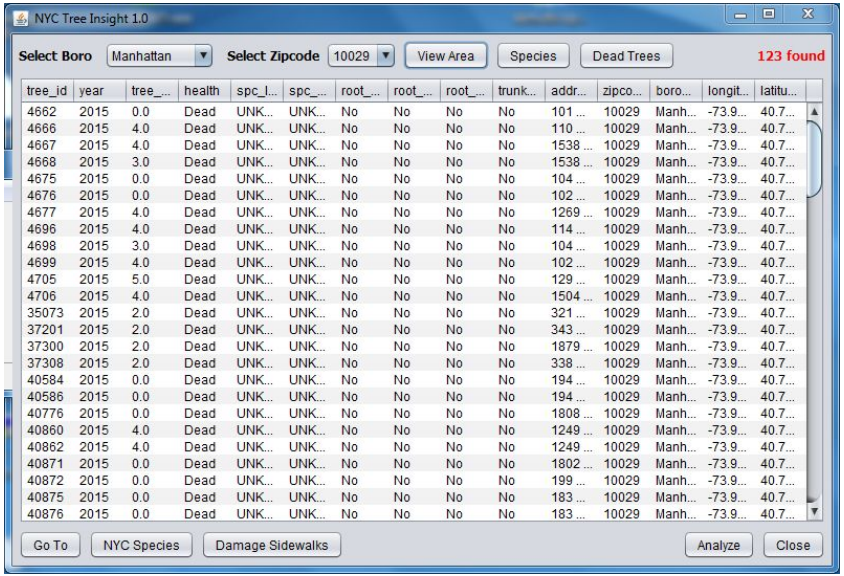
Variable Importance

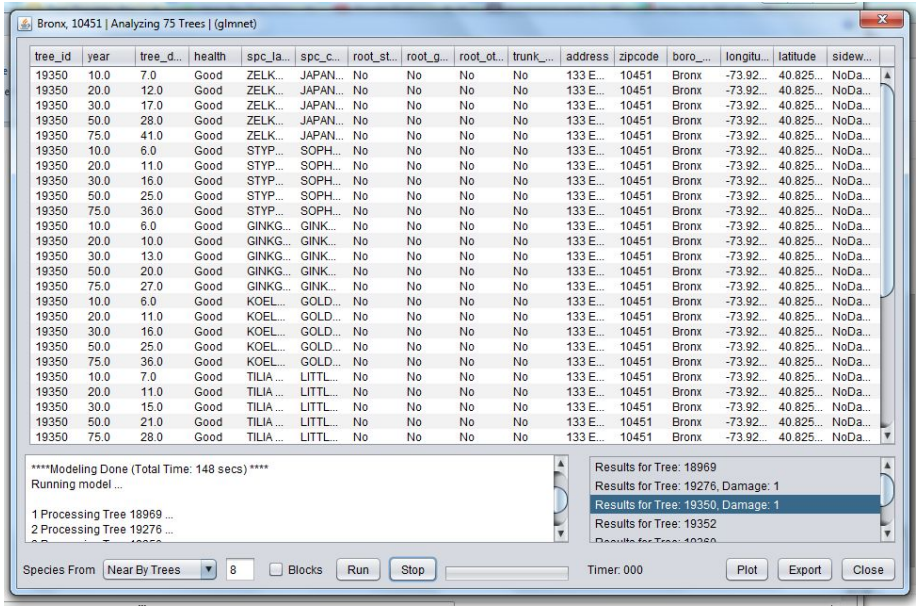




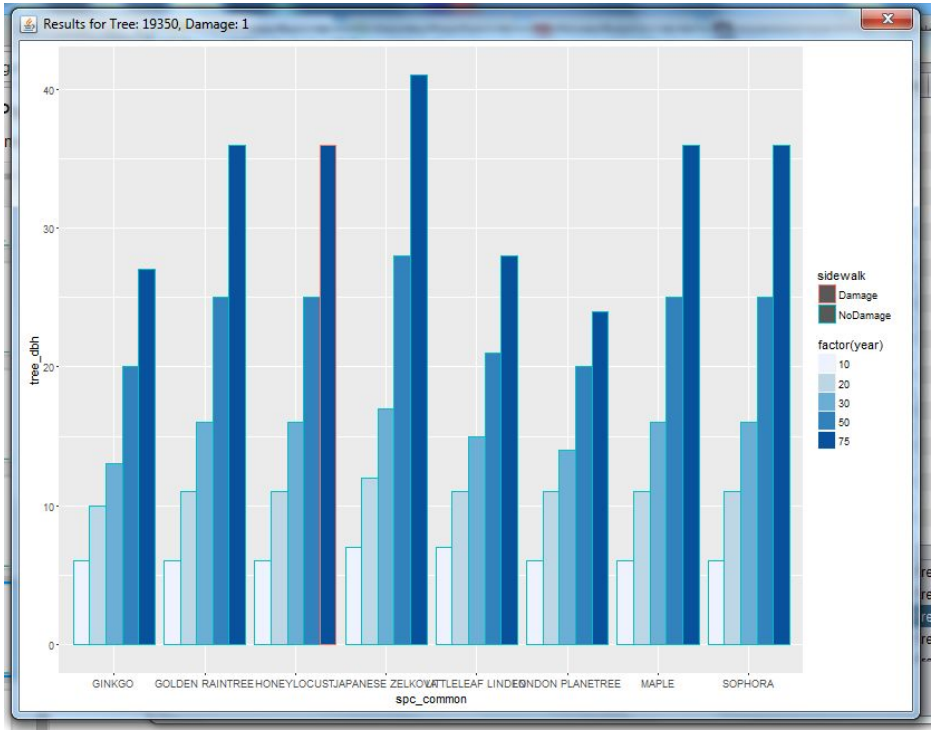
Desktop Analysis App “NYC Tree Insights”

Performs analysis on “dead trees” data to predict the potential for sidewalk damage at various years (10, 20, 30, 50, 75) in the future.





Visual Analysis



Next Steps

- Improve User Interface of Desktop Application
- Enhance Performance of Prediction Backend
- Migrate ML Backend to Amazon ML Services
- Create Web Application
- Create Mobile Application