NYC Environmental Analysis

Insight from Health Condition of Trees

Author:

Nian Yi (yn2336)

Abstract

In this project, I construct exploratory analysis, feature engineer, model evaluation as well as the importance feature analysis. I tried different models to classify and predict the health conditions of NYC trees. By comparing model with RMSLE (Root Mean Squared Logarithmic Error) and accuracy, I chose a model perform the best and then analyzed some of the critical factors that influence on environment.

Introduction and Motivation

NYC has over six billions of trees that been recorded including their health conditions (poor, fair, good), locations, tree diameters, whether they have a guard and etc. The health condition of trees actually indicate the environment condition around them. Therefore, by analyzing the health conditions of trees, we may answer the question like: which area in NYC has the best environment?

Besides this, I am also interested in what type of tree tend to live better in NYC area. Moreover, by searching for the factor that contributing to the health conditions, I hope to find a way to improve the health condition of trees and this can be helpful to the environment.

Methodology

1. An glance of the data and EDA

Our data does not have too much missing values or outliers. I check the variables including longitude, latitude, tree diameters and so on. There are no outliers for our geometric indicator such as longitude. Some of the tree diameters are zeroes and that means dead or stump. I just exclude them because it make no sense of the trees' "health". Also outliers including 0 at zip code are also excluded.

The overall condition of trees are distributed not very balanced. So I will use some of the model that are not influenced by this too much, for example the logistic regression and decision tree model.

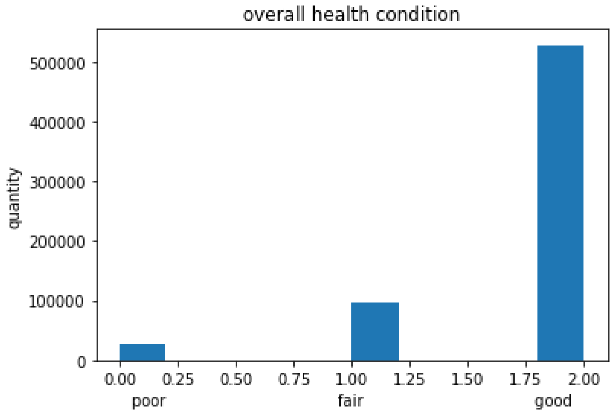


Figure 0

Next, I look at several features that might have influence on tree health: tree diameters, brunch problems, longitude, specific years and the tree types and etc.

Tree Diameter

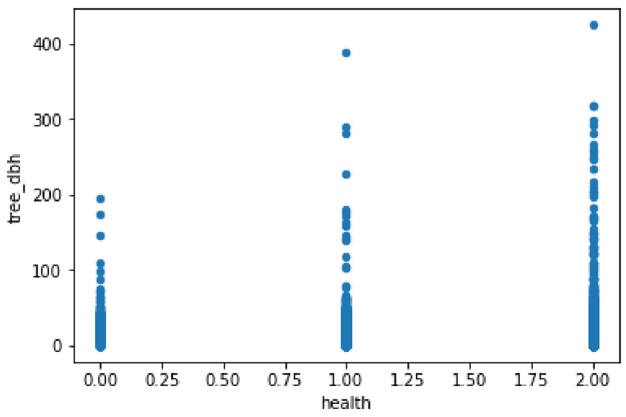


Figure 1

Health Condition for Trees have Brunch Problem

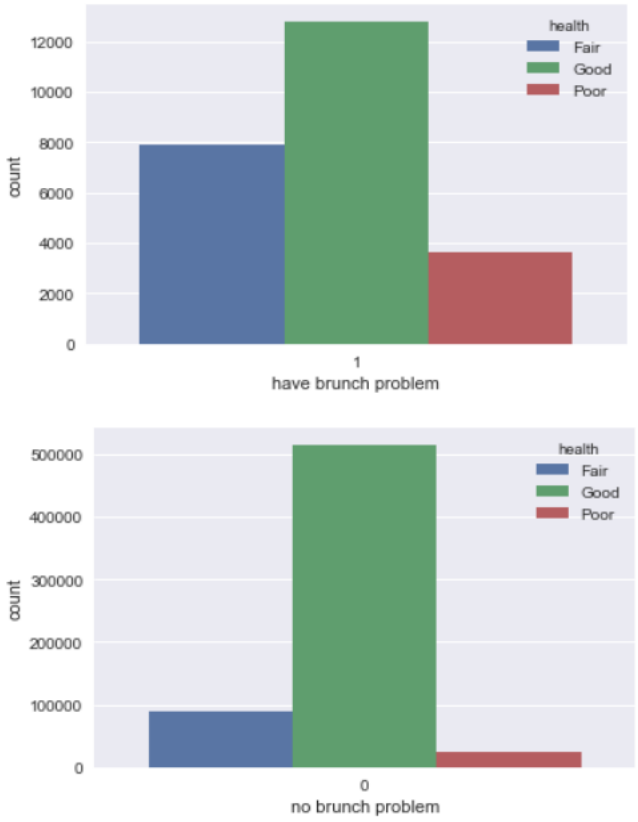


Figure 2

Geometric Influence on Tree Health

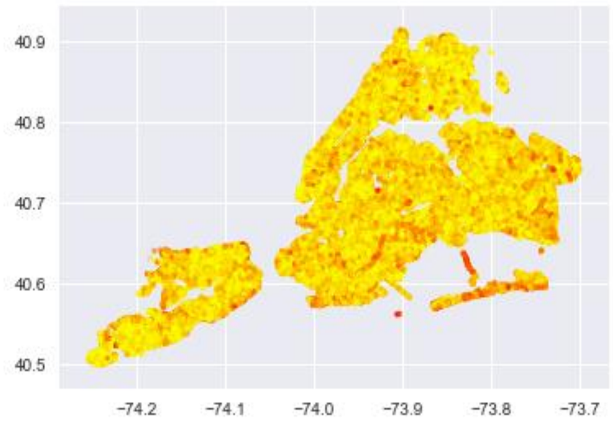


Figure 3

Difference among Tree Types in Their Health

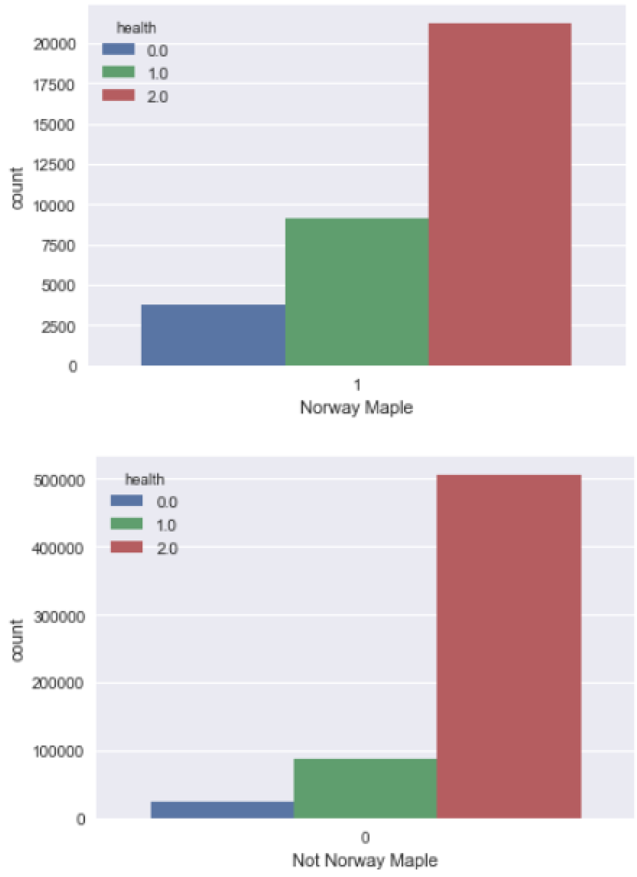


Figure 4

Health Condition for Trees Recorded in Different Years

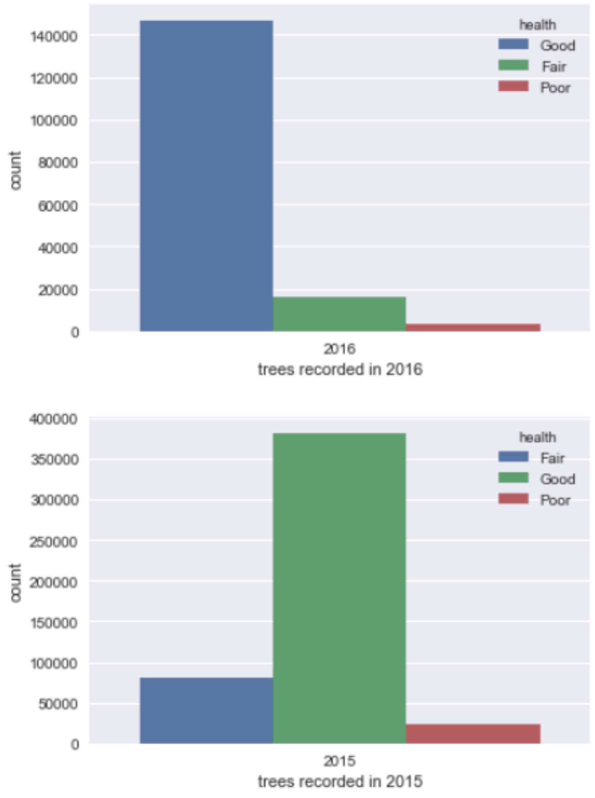


Figure 5

Findings:

1. Trees with bigger diameters seems healthier than smaller trees.
2. Trees' health are different if they have brunch problems compared to not.
3. Trees from some specific locations grows better than others.
4. Some specific tree types are healthier or unhealthier than others and we need to analysis which types they are.
5. Trees recorded in 2015 seems different to 2016, but the difference is not too much.

2. Transforming Data for Models

A big part of my project is to transform categorical data in a proper way so that can be used in my models. Almost all of the features are categorical, I go over them one by one and transform then using one hot encoding method.

Model

I will use 80%-20% training and testing split on all models. Within the 80%, I will use 5 folds cross validation and root mean square log error for parameter tuning.

I will use unsupervised learning approach such as K-means and PCA to transform data or reduce our dimensions of data. And then building model in a supervised way: logistic regression, ensemble of decision trees or KNN.

3. K-means with Logistic Regression

I first add 500 clusters derived by K-means with the latitude and longitude provided to substitute them. Here is a visualization of my clusters. It will mainly be used in logistic regression.

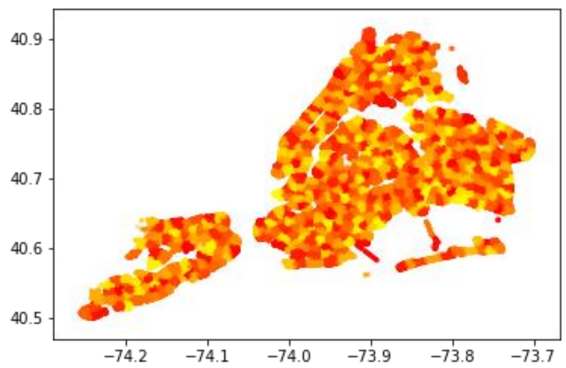


Figure 6

After adding clusters, I finally tuned a logistic regression model with L2 regularization parameters(C=1000, $\lambda = 1/1000$). Here is a feature importance of the model. We can see that the type of trees, the location of trees and the month that the tree was recorded have larger influence.

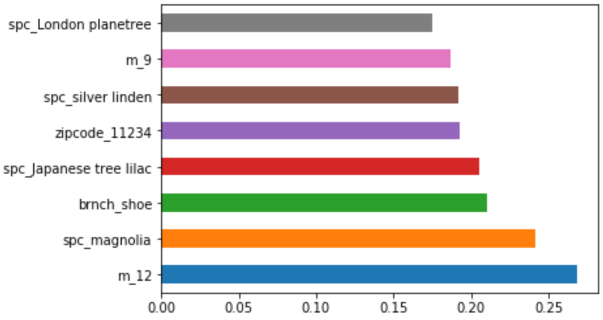
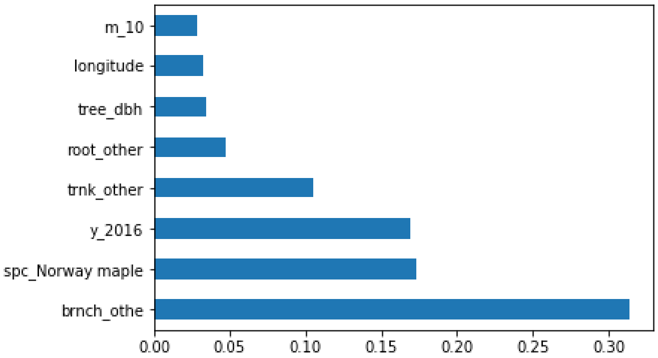


Figure 7

4. Building Decision Tree and ensemble method

The next method I tried is the decision tree. This method is relatively faster compared to SVM, especially when we have more features. I have several parameters to tune which includes max depth, sample leaf and sample split ratio. parameters of max depth.

Figure 8



Tuning process was verbose and I delete some of my trial in the code as well. Here is an example of finding tuning parameters for single parameter case (max_depth).

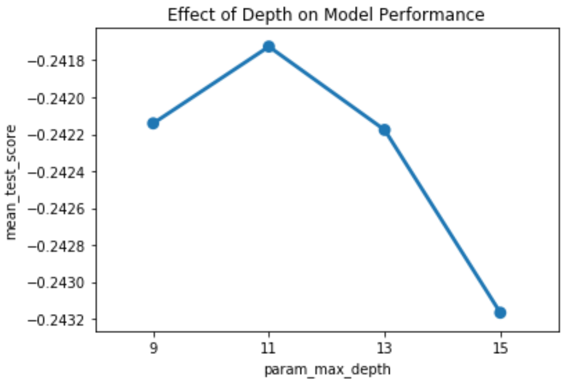


Figure 9

My best performance parameters are: Max depth: 13, min sample leaf: 4, sample split 0.1. RMSLE: 0.272 on test data.

I improved this model a little by using ensemble method that use my trees as weak learner. Here, 50 trees will result in a RMSLE of 0.264. Here is the feature importance of ensemble method and decision tree model (Figure 8). We can see that the longitude is a critical part for determining the health condition. Also, branch problem is the most harmful for trees.

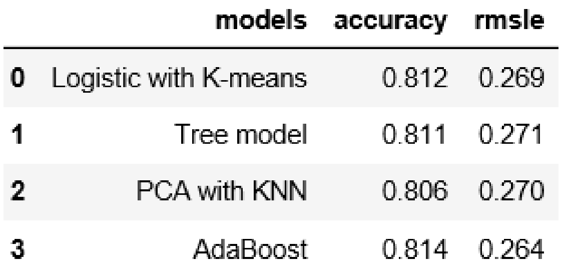
5. PCA with KNN Classifier

Curse of dimensionality will happen in our case since there are too many features after transforming categorical variable into dummy variables. So I decide to use PCA to reduce features to only 20 of them.

To build my KNN model, I choose my tuning parameters to be the number of neighbors and also the distance metrics. I finally get my best model to be a KNN with N=11 with Euclidean distance. Using the same test set, I got a RMSLE of 0.27 here.

6. Model Comparison

Comparisons are basically base on their accuracy rate as well as RMSLE. Here, the ensemble method or AdaBoost method performs the best.



	models	accuracy	rmsle
0	Logistic with K-means	0.812	0.269
1	Tree model	0.811	0.271
2	PCA with KNN	0.806	0.270
3	AdaBoost	0.814	0.264

Figure 11

Analysis

Ensemble method is the best of my models and from the feature importance, we know that the branch problems (other) seems to be the most important factor that influence on the tree health. Others are the tree type (Norway maple), trees recorded in 2016, trunk and root problems, tree diameters, longitude and etc. And from the result, we may infer that:

1. Branch problems are more severe than root and trunk problems for trees.

2. We might need to dig in more to find the reason that why tree from 2016 are healthier than other years. Future works may include weather data.

3. Norway Maple can grow better or worse than other trees.

4. Longitude or the location is a factor that influence tree health. We may want to know which area is best: for example, in logistic regression, we know that the tree that living in Four Sparrow Marsh Park (zip code 11234) at Brooklyn is the best. We can infer that the environment is better as well.

The decision tree model’s feature importance is not stable. So we may also refer to the importance provided by the logistic regression model. From the results from logistic regression, Japanese tree and silver linden also have better health. According to the data, the proportion of trees in good health conditions among Japanese tree, silver linden are 4 percent higher than others. For Norway maple, it is 20 percent worse than others. This means that, without Norway maple, 82% percent trees are healthy. However, only 62% percent are healthy including Norway maple.

Conclusion

In conclusion, all of models provide a similar accuracy. They are not too high and there might be some inaccuracy in determining the health condition of trees. Also, we might need more data for model improvement. But based on my current models, I can deduce that branch problems are a major problems among NYC trees. Also, some trees are easy to grow in NYC area like London planet tree or Japanese Tree while some are not like Norway maple. Norway maple is extremely different from others and I research on this trees more: It was imported from Europe and invasive and aggressive to other trees. Now NYC has more than30 thousands such trees. They are "hated" by people living in urban area since they destroy pavement and require expensive repairs.

Figure 12

What about the environment?

The environment might become worse in 2016 compared to 2015 since trees seems not healthy in 2016. This might because of the weather, temperature and so on. We need more information in determining this relation. Overall, there is no extreme different between locations by observing the condition for trees. Location like Sparrow Marsh are better for trees to grow and may have better environment.