Machine Learning for Environmental Researchers: Plant Classification

By

Trist'n Joseph – trisxcjoseph@gmail.com

December 2019

‘Classification trees’ can be used to predict the membership of objects in a specific categorical variable based on the measurements on one or more predictor variables. The idea behind this type of analysis is to predict the group membership of various objects based on the data provided for the objects’ variables. There are two assumptions made for this system of classification;

1. There exist multiple groups within the data set and,
2. Observations are known to be members of only one group.

The latter assumption means that if the groups in the data were known to be Male and Female then one observation cannot belong to both Male and Female. Thus, based on the data, would it be possible to determine a specific object’s class? This type of analysis is useful because it can provide an efficient and robust system for automating tasks, such as sorting coins, once the analysis is done correctly.

In this project, I use discrimination analysis to attempt to classify types of physical trees based on their characteristics within the data set. This study is relevant because trees contribute to the environment by providing oxygen, preserving soil, providing an ecosystem for wild animals and reducing the intensity of the greenhouse effect. It is also known that a tree’s chance of survival depends on factors such as climate, exposure to sunlight, location and soil type. If it is possible to classify trees based on observable patterns within the data, then it might assist arborists determine characteristics of particular ecosystems.

The ‘2015 Street Tree Census’ data set contains approximately 684,000 rows and 45 columns, where each row refers to a particular tree and each column refers to the variables captured within the data set. The assumptions of discrimination analysis are met within the data set as the species of the tress is provided and each observation belongs to only one specie. Upon initial analysis, it was found that the set contained at least 133 species of trees. However, I realized that multiple species of trees belonged to similar families of trees. Therefore, I decided to group trees into their respective family type and conduct analysis on the family type rather than on the species of the tree.

The data set contains at least 28 different family types which was consolidated into 11 categories. I chose to only use 11 categories because many categories accounted for a very minimal portion (less than 1%) of the data set and many trees were unnamed within the data set. Therefore, I chose to explicitly name the top 10 family types with the highest proportion within the data set and group all other trees as “Other”. Table 1 shows the distribution of trees within their family types.

Family Name	Totals	Proportions
Other	114615	0.1676
Rose	106177	0.1553
Legume	88913	0.1300
Maple	88739	0.1298
Sycamore	87014	0.1273
Beech	84650	0.1238
Elm	44173	0.0646
Olive	25245	0.0369
Maidenhair	21024	0.0307
Spurge	12293	0.0180
Sourgum	10945	0.0160

Table 1: Proportion of Tree Family Within the Data Set

Figure 1 shows the scatter plots of the variables within the data set using colours to segregate the observations based on their family type. Based on this figure, there is no clear relationship between any two variables along with the specific family type. Additionally, based on the number of family types, comparisons of visual representations proved to be extremely difficult. Therefore, I chose variables such as the tree's size since I assume that it would be related to the tree's family type. I then created multinomial logistic regression models to determine the probability of a tree's membership within a family.

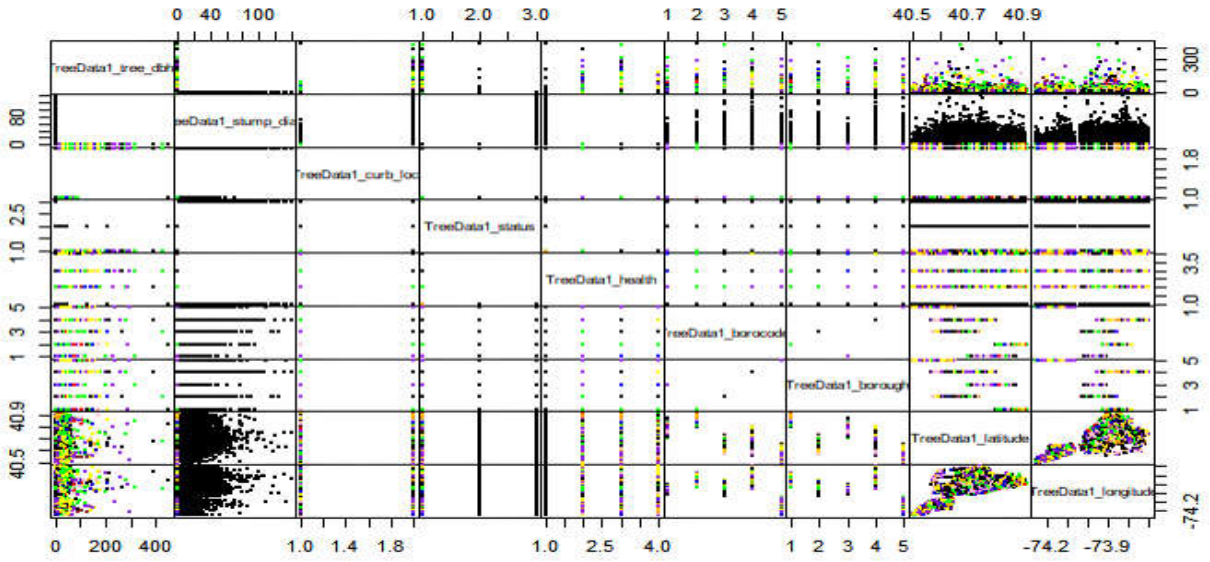

Figure 1: Pairs Plot Showing the Relationship between Variables

The multinomial logistic regression requires a reference class, and its output predicts the probabilities that a particular observation belongs to the other classes based on the information compared to the reference class. I chose the Rose family to be the reference class in my model because I wanted to compare the rest of trees to the known family which had the largest proportion within the data set. Based on the results from Table 1, Other could have been the reference class. However, I chose not to use Other because it consisted of many different family types with a large variation between types, which could reduce the reliability of my model.

The final model used for analysis was as follows (Equation 1);

$$\ln\left[\frac{P(Y = category_i|X)}{P(Y = Rose|X)}\right] = \beta_0 + (\beta_1 * TreeDiameter) + (\beta_2 * TreeStatus) + (\beta_3 * TreeHealth) + (\beta_4 * TreeStreetLocation) + (\beta_5 * TreeBorough) + (\beta_6 * TreeLatitude) + (\beta_7 * TreeLongitude) + \varepsilon_i$$

Variable Name	**Description**
TreeDiameter	A numerical variable which captures the diameter of the tree, measured approximately 137cm above the ground.
TreeStatus	A categorical variable (factor) which indicates whether the tree is Alive, Standing Dead, or a Stump.
TreeHealth	A categorical variable (factor) which indicates the surveyor's perception of the tree's health. Within the data set, these are coded as Unknown, Fair, Good, and Poor.
TreeStreetLocation	A categorical variable (factor) which provides the location of the tree bed in relation to the street curb. Trees are either along the curb or offset from the curb.
TreeBorough	A categorical variable (factor) which provides the NYC Borough where the tree is located. Trees are located in The Bronx, Brooklyn, Manhattan, Queens, or Staten Island.
TreeLatitude	A numerical variable which provides the latitude location of the tree in decimal degrees.
TreeLongitude	A numerical variable which provides the longitude location of the tree in decimal degrees.

Table 2: Description of Variables in Multinomial Logistic Regression

Table 2 provides the descriptions of every variable within the model. The output of Equation 1 produces ten models which would be compared to the reference category, each with their own set of coefficients. Table 3 shows the summary of model coefficients for the first two categories.

Category	Intercept	Tree Diameter	Tree Status	Tree Health	Tree Street Location	Tree Borough	Tree Latitude	Tree Longitude
Beech	-125.052	0.154	-0.8990	0.006	-0.797	-0.209	2.745	-0.199
Elm	-34.252	0.051	-1.0470	0.038	-0.658	-0.388	1.363	0.258

Table 3:Summary of Coefficients (first two categories)

The coefficients provide a sound interpretation of the model, and the likelihood for a tree to be a member of a certain family type. For the Beech family; compared to the Rose family, the logit (which is the logarithm of the odds) increases by 0.154 for every unit increase in the tree's diameter holding all other variables constant. Whereas for the Elm family; compared to the Rose family, the logit increases by 0.051 for every unit increase in the tree's diameter holding all other variables constant.

The performance of the model is based on how well it classifies trees into their correct family types, since the family types are known in this data set. The model performed best when classifying family types which had more than 10% representation within the data set. If a tree was known to be in a category with fewer than 10% representation, such as Olive, the model incorrectly classified the tree as Other or Legume. Table 4 provides a sample of the model's classifications. Here, classification was chosen by creating a table of fitted probability values of every family type for all the trees within the data set. Then, the highest family type probability was chosen for each tree. Based on the information provided in Table 4, it seems like the model performed ideal 60% of the time. This could be considered as 'good' performance since it is accurate more than half of the time, but this model is far from ideal.

Model Family Type	Known Family Type	Highest Probability
Beech	Beech	0.239
Sycamore	Beech	0.400
Other	Other	0.211
Other	Other	0.212
Other	Other	0.201
Legume	Spurge	0.232
Maple	Maple	0.228
Legume	Elm	0.226
Legume	Olive	0.224
Rose	Rose	0.225
Legume	Legume	0.217
Other	Maidenhair	0.229
Legume	Sourgum	0.239
Sycamore	Sycamore	0.249
Maple	Maple	0.204

Table 4: Sample Model Performance

I hypothesized that the model struggled with correct classification due to the lumpsum of trees grouped together in the Other family. It must be noted that the further segregation of trees does not automatically mean that the model would perform ideally 90% of the time, but it might have decreased the amount of times a tree was incorrectly classified as Other. The model also struggled due to a lack of variables which could be closely related to the tree's family type, such as the tree's height, the canopy width, or even the type of leaf.

To test my hypothesis, I decided to remove all observations which fell into the Other family and redo my analysis based on all the remaining classes. Table 5 gives the summary of the model coefficients. These should be interpreted similarly to the results of Table 3. As can be seen from Table 5, the model coefficients are similar to those of Table 3, with the exception of the intercept Tree Status. For the analysis, I kept the Rose family as my reference class since it would still be the family type with the highest proportion within the data set.

Category	Intercept	Tree Diameter	Tree Status	Tree Health	Tree Street Location	Tree Borough	Tree Latitude	Tree Longitude
Beech	-59.004	0.159	-59.251	0.0027	-0.805	-0.214	2.631	-0.157
Elm	-16.259	0.054	-16.351	0.0339	-0.656	-0.389	1.278	0.248
Legume	-135.652	0.084	47.587	-0.0749	-0.448	-0.304	2.421	0.125
Maidenhair	-74.659	0.052	-74.677	0.1214	0.034	-0.263	3.481	-0.081
Maple	73.849	0.148	42.457	-0.2143	-0.092	0.045	-0.641	1.234
Olive	82.021	0.115	40.461	-0.1254	0.147	-0.123	-0.489	1.415
Sourgum	-66.131	0.109	-66.183	0.0518	-0.897	-0.002	-0.889	-2.255
Spurge	88.694	-0.179	47.577	-0.0181	0.261	0.018	-0.437	1.625
Sycamore	112.141	0.254	43.491	-0.0862	-0.713	-0.445	-2.496	0.737

Table 5: Summary of Coefficients of Model Excluding 'Other'

Table 6 gives a sample of the new model's performance. Unfortunately, the model incorrectly classified most observations as Legume. I found this surprising because of my previous assumption regarding the Other family. However, it is possible that the model performed better previously since many observations had a great chance of being classified as Other; including observations which were known to be in the Other family. Thus, my new hypothesis is that the model would perform better if the data set including more quantitative variables related to specific tree families, such as the average height. This is because the model will highlight the differences between tress if a certain tree family is known to have a higher average height compared to another tree family.

Model Family Type	Known Family Type	Highest Probability
Sycamore	Beech	0.429
Sycamore	Beech	0.304
Legume	Spruge	0.257
Legume	Spruge	0.289
Legume	Elm	0.292
Legume	Elm	0.293
Elm	Elm	0.279
Beech	Sycamore	0.251
Sycamore	Sycamore	0.258
Legume	Rose	0.280
Rose	Rose	0.280
Rose	Rose	0.280
Legume	Maidenhair	0.254
Legume	Olive	0.280
Legume	Sourgum	0.221
Legume	Spurge	0.300
Legume	Maple	0.281

Table 6: Sample Model Performance Excluding 'Other'

Therefore, although this study did not provide a rigorous model for classifying trees into their respective categories, it showed that it is possible to develop a classification system for specific trees based on a few variables regarding the trees in question. As stated before, this study can be improved by capturing quantitative variables which are more unique to a tree's family type.