Forest Cover Analysis

TCSS-551 Term Project

Jonathan McFadden

Introduction

The goal of this project is twofold. First, we seek to create an accurate big-data/machine-learning model to predict which type of trees will grow best in a certain area based on the conditions of that area. Finally, we seek to determine which features of a given area are most important for determine which type of tree will grow best there. To do this, we have chosen three different methods:

- Linear/Logistic Regression
- Decision Trees
- Random Forest

We chose to try a Linear/Logistic regression first because they're provide a simple and low computational cost way to see a relationship between the features and outcomes of the provided data does indeed exist. This method will be unlikely produce predictions with good accuracy due to its simplicity, however should these methods show a relationship or relationships within the provided data, that would be a positive indicator to proceed to more complicated and compute intensive algorithms.

METHOD A - Linear/Logistic Regression

As described above, the initial method we tried were Linear/Logistic Regressions. We tried three variations on this method. First, we ran a simple linear regression. This gave a poor result, but *did* indicate that there was a relation between the features and outcomes in the data. This linear model gave an R^2 value of 0.4214 on the test data and 0.4006 on the training data. This is odd as, one would expect this value to be higher for the training data than the test data. This trend also continued when comparing via MSE[1], as the training data had an MSE of 2.3952 while the test data had an MSE of 2.3257.

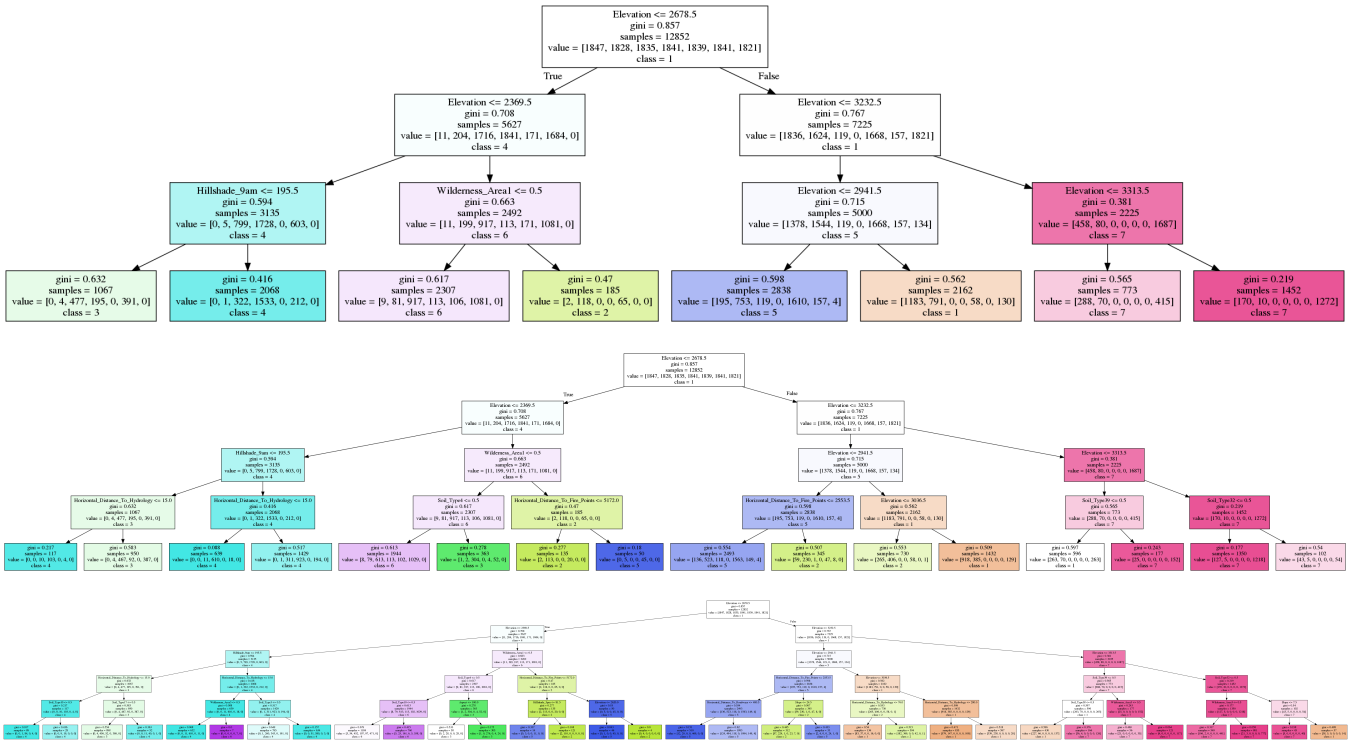
We followed this basic Linear Regression model with two Logistic Regression models; one based on fitting a binary problem for each class (*OVR*) and the other based on a multinomial fitting (*MULTNOM*). The *OVR* Logistic Regression had a mean error of 0.6711 and an MSE of 2.8735 on the test data. The *MULTNOM* Logistic Regression did slightly better on mean error with 0.6653 and slightly worse on the MSE with 3.2866 on the test data. Interestingly, the *OVR* Logistic Regression had a better MSE on the test data than on the training data (2.9991) while the *MULTNOM* Logistic Regression had a better mean error on the test data than the training data (0.6721).

While this method did establish that *there is* a relationship between the features and outcomes in the data, it does not have very good accuracy. More importantly, it does not tell us anything about which features are most important for determining which type of tree will do best in a certain area. However, now that we know there is a relationship within the data, we can move on to more complicated and computationally intense methods to determine the importance of the various features.

METHOD B - Decision Trees

The first detailed method we chose to use was Decision Trees. We chose to use this method because it is relatively simple[2], yet very powerful as it has the ability to give us an idea of which features are most important. To determine feature importance, we built three different decision trees, with 3, 4, and 5 levels respectively. These decision tress can be graphically represented as

[1] mean squared error
[2] as far as ML algorithms go



From these trees, we can clearly see that **Elevation** is the most important determining feature, by far. We can also see that **Elevation** is flowed by, in no particular order, the features:

- **Horizontal Distance To Fire Points**
- **Horizontal Distance To Hydrology**
- **Hillside Shade 9am**
- **Wildernes Area 1**
- **Soil Type4**

For the last part of applying the Decision Tree algorithm to this data, we constructed a Decision Tree with no limit on its depth to use for making predictions. This complete Decision Tree model has a mean accuracy of 0.8078 and an MSE of 1.7571, beating the Linear/Logistic Regression by a *large* margin.

METHOD C - Random Forest