Analysis of Key Factors for Forest Type Prediction

Michael Remington, *Aspiring Ecologist*

09/01/2019

Abstract

The purpose of this study is identifying the most important factors in predicting the predominant tree species in a 30 x 30 meter plot. The U.S. Forest Service produced the plots and the UCI machine learning repository hosts the data set of 15120 observations. Forest Cover Type is split into 7 different tree species and the data set contains 11 climate/geographic variables [1]. To predict the Forest Cover Type and gauge the relationships of independent variables collected, we use a random forest classification model. Random forests work well with categorical data and gives insights on the importance of the measures used for predicting [2]. Our final model was able to assess our predictor variable correctly at a 94.3 percent accuracy rate and determine the strongest relating vectors: Elevation, Soil Label, Horizontal Distance to Roadways, Aspect, Horizontal Distance to Hydrology, Vertical Distance to Hydrology, and finally Slope.

Contents

1. Introduction

For this project provided by the learning website Kaggle, the U.S. Forest Service and US Geological Survey provided a raw data set from the Roosevelt National Forest of Northern Colorado to then allow competitors to build a machine learning model to predict Forest Cover Type. These types of models can be useful to help determine important environmental characteristics of certain climates, reduce sampling costs for future experiments, and can be applied to other forests of similar geography and Forest Cover Types.

Observations in the data are a $30m$ x $30m$ experimental plot, with 15120 plots in total. Each point provides data on; 10 continuous environmental factors, 4 binary factors displaying the wilderness region, 40 binary factors displaying the Soil Type, and finally, the data set contains a predictor variable titled Cover Type which contains 7 different predominant tree species. This is the dominant tree population in the plot determined by the forest service. These are evenly distributed (2160) observations for each cover type, the species of tree is listed below in list 1.

Prediction Classes

- Spruce/Fir
- Lodgepole Pine
- Ponderosa Pine
- Cottonwood/Willow
- Aspen
- Douglas-fir

- Krummholz

Our goal is to explore the relationships between these variables and build a model to determine which ones are significant and if this data set allows us to predict the Forest Cover Type. The independent variables we have to work with are the plot's: Elevation, Aspect, Slope, Horizontal Distance To Hydrology, Vertical Distance to Hydrology, Horizontal Distance to Roadways in meters, Hill Shade at 9am in summer solstice measured with a 0-255 index, Hill Shade at 3pm in summer solstice measured with a 0-255 index, Horizontal Distance to Fire points in meters, Wilderness area which is 4 separate geographic regions of the state park, and soil type of 40 different categorical values of the type of soil that the plot has. This includes the family of soil, the texture, and if it is a mix of other types. For the purpose of this study we treat this complex as a discrete variable with 40 different values.

2. Exploratory Data Analysis

Before we start building the model, we need to get more familiar with the relationships between our predictor variable, and our 11 independent variables to view the dynamics between these variables and the spread of data. This is important so we don't misinterpret relationships and avoid overfitting the model we will build. This can also help distinguish the importance of certain variables compared to others.

In the Figure 1, we can look at the dynamic relationships between our continuous variables. Nothing in this graph inherently describes which of these variables would be good at predicting Cover Type, however, we can still learn a lot from it. We can see Aspect has a close relationship with Hill Shade at 9am with a R^2 value of -0.59 which is a somewhat strong negative correlation and an even stronger relationship with Hill Shade at 3pm with a R^2 value of 0.64. This makes sense as Aspect is the measure of the compass direction of the slope which would influence how shade is distributed on a slope. We also see a clear relationship in these variables in the odd data density distributions in the lower triangle of the matrix. Our goal when composing independent variables for a predictive model is to avoid colinearity as it can compromise the model.

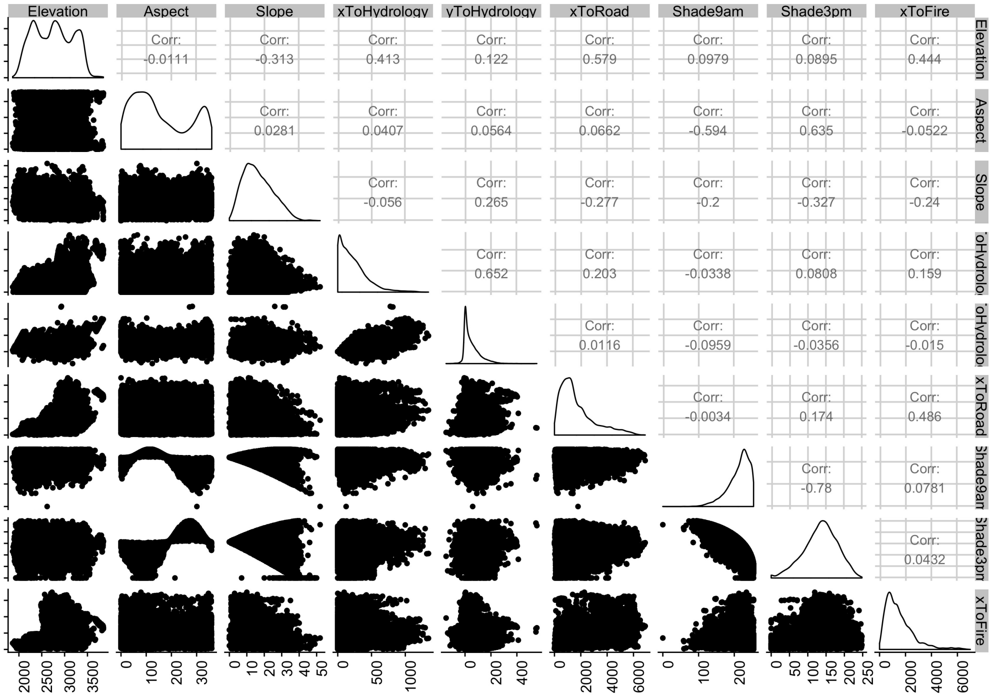

Figure 1: Plot matrix of our independent variables.

Now that we have looked at our continuous variables, we need to further examine the discrete variables of Wilderness Type, and Soil Type. In Figure 2, we can visualize how many plots were taken in each of the 4 wilderness areas. From this graph, we can see an uneven sampling distribution for our data set. This could have bias affects on our model if this variable comes up as an important vector.

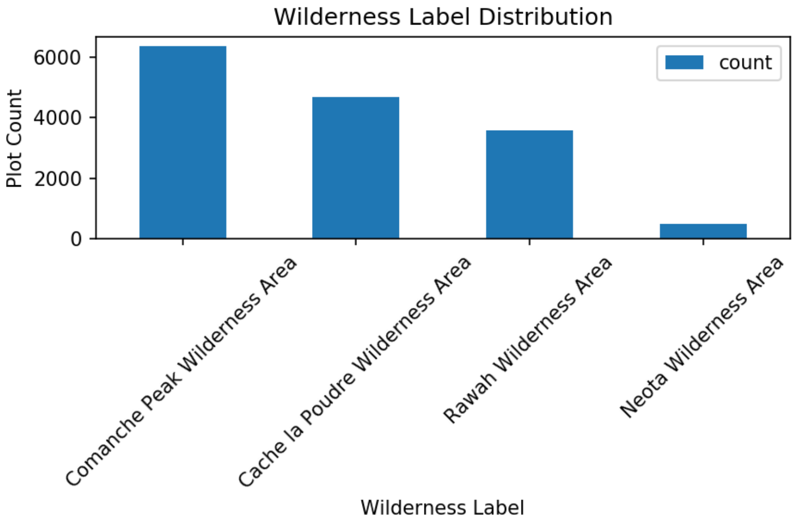

Figure 2: Bar chart of the number of plots in the four geographic areas.

For our last independent variable, Soil Type, we see in the heat map shown in Figure 3 that there is a visual relationship between certain Forest Cover Types and Soil Types. Douglas Fir, Cottonwood/Willow, Krummholz, and Ponderosa pines seem to prefer certain types of soil while Aspen, Lodgepole Pine, and Spruce/Fir tend to be more flexible. More analysis on this would be necessary to find direct relationships, but this is a good indicator that Soil Type has a relationship with Tree Cover Type.

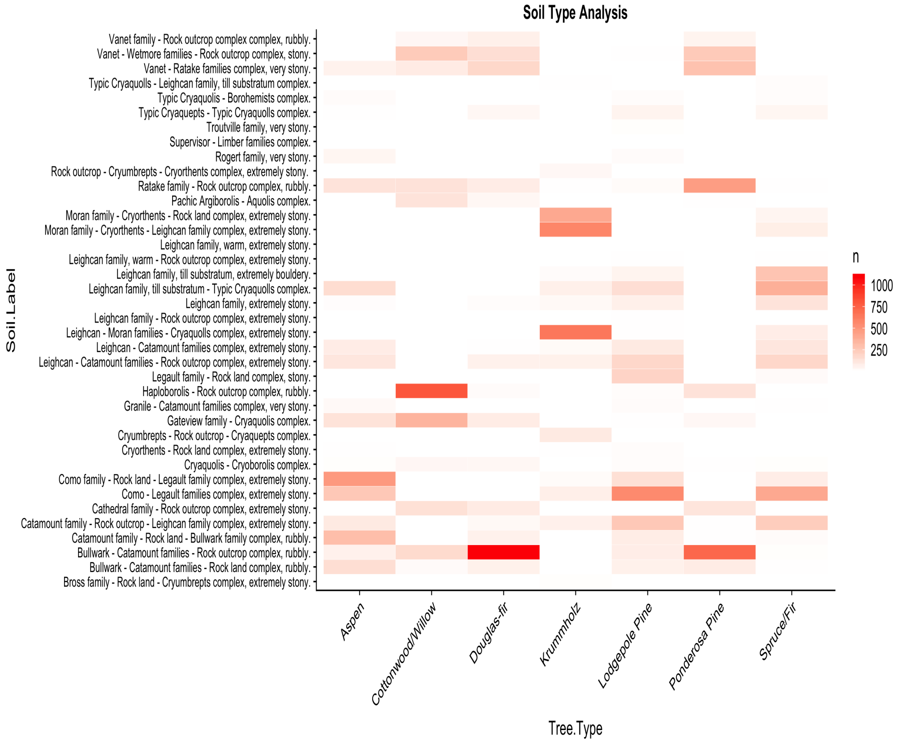

Figure 3: A heatmap displaying the density of plots which have certain Soil Type characteristics.

3. Methods and Model Building

Now that we have surface level view of the relationships between our vectors, we can now build a model to determine which factors are the most important in predicting Forest Cover Type and which factors we may want to leave out of sampling next time around. For this model we will split our data into a training and testing set and use the training set generate a random forest classification model.

We went with this model because it is a classification model that works well with categorical data computationally and provides many insights on the relationships of the predictor variables [2]. We will generate two models, one with all the variables, and the second with colinear variables and forest characteristics specific to the Roosevelt National Forest removed. To test the accuracy of the model, we will split our data set into two separate sets, a random 2/3 of the data will go to the training set and 1/3 will go to our testing set. We will then try to predict the Cover Type in the testing set and measure how many we are accurately able to estimate. To measure the importance of each independent variable, we will use the mean decreasing gini score which is a continuous metric which quantifies weight of how important a variable [2]. A large mean decreasing gini score means the variable was important for predictive capability.

4. Results and Discussion

4.1. Model 1

After generating our training model the out of bag error rate, or the accuracy rate the model generates against itself showcases an 86.1 percent predictive accuracy. When we use our model to predict the values in the testing set, we are able to predict the correct Forest Cover Type with a 95.3 accuracy rate out of a sample size of 5040 plots.

To further diagnose the validate of our model, Figure 4 showcases flaws in our model. We see we are unable to correctly predict Spruce/Firs with an error rate of 7.6 percent and Lodgepole Pines have an error rate of 11.7 percent. The largest trend in miss diagnosing the testing set's Cover Type is is false positively predicting Spruce/Firs instead of Lodgepole Pines. This means we may need to figure out a variable that may distinguish these two Cover Types more.

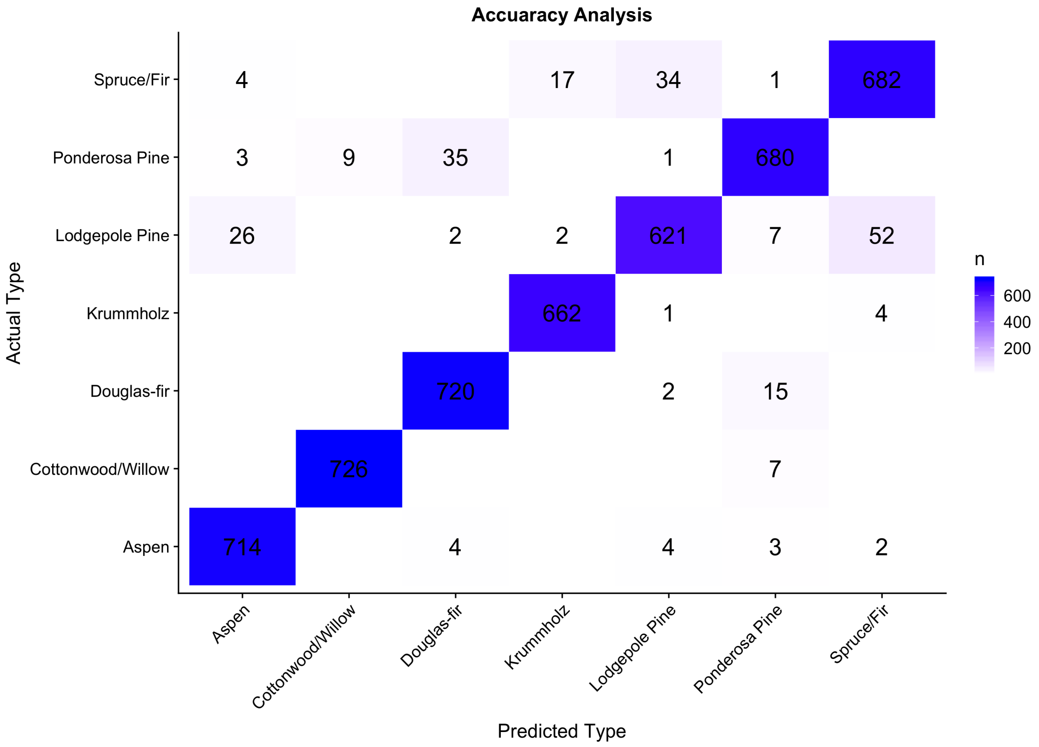

Figure 4: Heat map showcasing how accurate our model is in predicting each of our variables.

Now that we have seen the accuracy of our initial model, we need to determine the importance of our independent variables to improve future sampling methods. Figure 5 showcases a clear dominance of predictive power using Elevation and Soil Type depicted by the mean decreasing gini metric described earlier. The rest of our variables have similar importance metrics so we see these two variables are the top drivers of our model and accuracy.

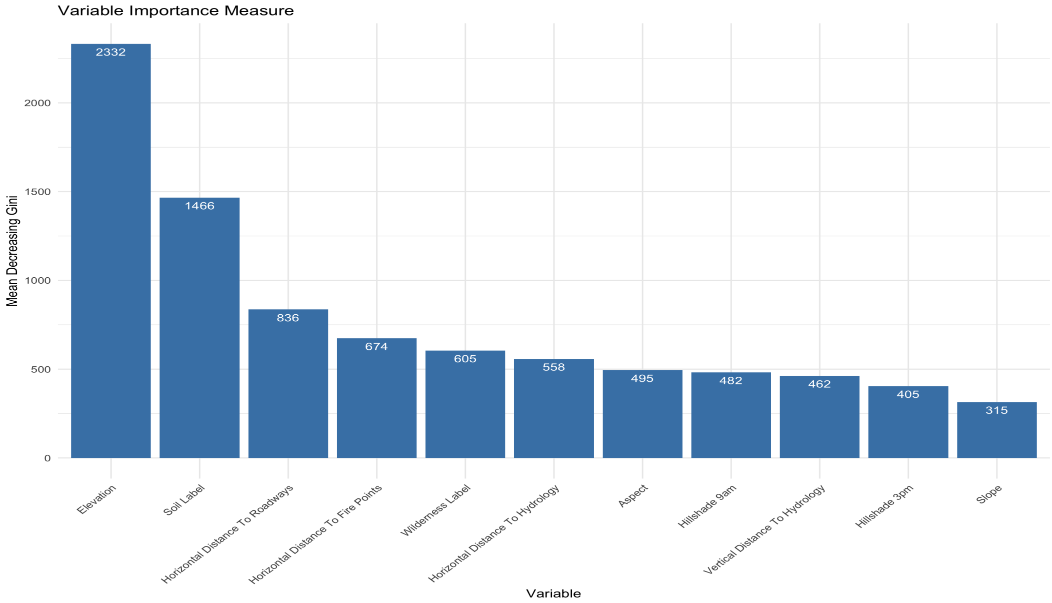

Figure 5: Bar chart quantifying the importance of our vectors.

4.2. Model 2

When analyzing Figure 1, we noticed some variables are significantly correlated which could lead to overfitting and colinearity, thus compromising our model. To fix this, we need to remove variables that are correlated. Also, it would be cost effective for future sampling and more applicable to be able to use this model for other geographic locations. So we removed the variables listed below in red 4.2. After removing said variables, we rerun the model and reevaluate whether we are able to predict Forest Cover Type against the testing set with sufficent accuracy.

Removed Variables

- Hill Shade 9am
- Hill Shade 3pm
- Horizontal Distance to Fire Points
- Wilderness Label

The new model has a predicting accuracy of 94.6 percent success rate. Compared to the first model with 11 independent variables, we lose less than 1 percent predicting accuracy and avoid overfitting with only 7 predictor variables. In Figure 6, we can see a side by side comparison in how accurate our model is in predicting each of the Forest Cover Types. The Cover Type that was negatively impacted the most from the loss in variables is the Douglas Fir as it incurred a drop in 2 percent accuracy loss. Other than that, the differences are minuscule, showcasing the variables we dropped weren't heavy drivers of our model.

Now that we have a trimmed model, the final step is to reassess variable

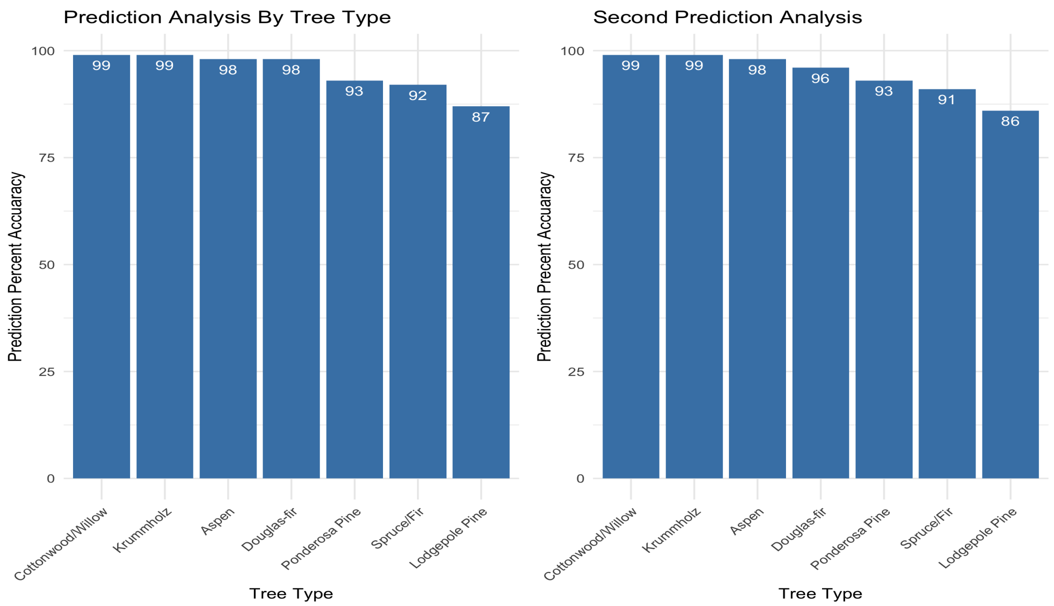

Figure 6: Bar chart of the accuracy of the first model vs the second model against the testing set.

importance. In Figure 7, we still see the same top 3 drivers of the model are Elevation, Soil Label, and the Horizontal Distance to Roadways. An interesting dynamic that has emerged is that Aspect more than doubled in importance once we removed Hill Shade variables. This is further evidence there was a statistical relationship between those Aspect and Hill Shade which makes sense given the nature of the measures. This showcases we have a reliable model to predict Cover Type with the variables: Elevation, Soil Label, Horizontal Distance to Roadways, Aspect, Horizontal Distance to Hydrology, Vertical Distance to Hydrology, and finally Slope.

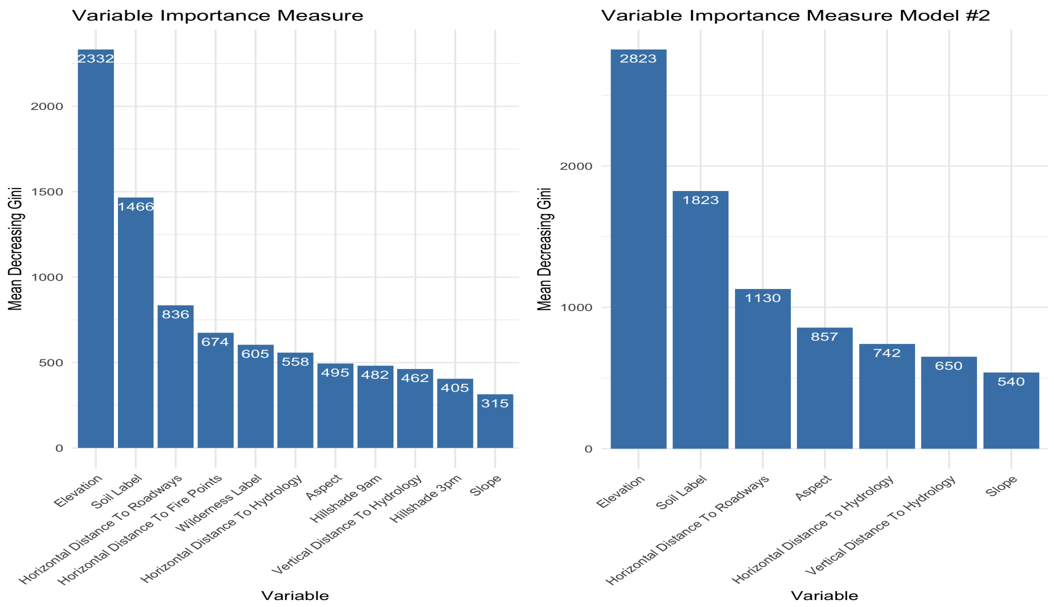

Figure 7: Bar charts of the accuracy of the first model versus the second model against the testing set.

5. Conclusion and Future Work

Although exhilarating and rewarding, environmental sampling is an expensive and arduous task so reducing the workload on the US Forest service is important. This model can be a tool for the U.S. Forest Service to reduce the amount of data they have to sample in similar climates and ecology to the Roosevelt National Forest of Northern Colorado. Our final model was able to predict Cover Type at a 94 percent accuracy rate using the 7 variables Elevation, Soil Label, Horizontal Distance to Roadways, Aspect, Horizontal Distance to Hydrology, Vertical Distance to Hydrology, and Slope.

Some future work that could be done with this report is to separate Soil Type into several variables. The Soil Type is a long string that describes the soil complex with high detail, but it can be split into several different sub-variables such as Soil Family, Soil Composition, and Complex Type. I think this would be beneficial in further extrapolating on the relationship between Soil and Forest Cover Type.

References

[1] M. Bache, K. Lichman. Uci machine learning repository, 2013.

[2] Josh Stramer. Statquest: Random forests part 1 - building, using and evaluating, 2016.